Überrelaxation bei linearen Gleichungssystemen mit schiefsymmetrischer Koeffizientenmatrix

Inaugural-Dissertation
zur Erlangung des Grades eines Doktors
der Naturwissenschaften
der
Mathematisch-naturwissenschaftlichen Fakultät
der
Eberhard-Karls-Universität
zu Tübingen

vorgelegt von

Wilhelm Niethammer
aus Unterjettingen (Württemberg)

1964

7719

1963

Überrelaxation bei linearen Gleichungssystemen mit schiefsymmetrischer Koeffizientenmatrix

Inaugural-Dissertation
zur Erlangung des Grades eines Doktors
der Naturwissenschaften
der
Mathematisch-naturwissenschaftlichen Fakultät
der
Eberhard-Karls-Universität
zu Tübingen

vorgelegt von

Wilhelm Niethammer

aus Unterjettingen (Württemberg)

1964

Tag der mündlichen Prüfung: 21.11.1963

Dekan: Prof. Dr. G. Möllenstedt

1. Berichterstatter: Prof. Dr. K. Zeller

2. Berichterstatter: Prof. Dr. H. Kneser

R. UNIV. BIBLIOTHEEK LEIDEN

Herrn Prof. Dr. Zeller möchte ich für die Anregung der Arbeit, für die reichlich gebotene Möglichkeit, die Rechenanlage des Tübinger Rechenzentrums zu benutzen, und für wertvolle Ratschläge herzlich danken. Ferner danke ich Herrn Doz. Dr. Ehlich für nützliche Hinweise bei der Niederschrift der Arbeit.

Digitized by the Internet Archive
in 2026

https://archive.org/details/IA41547601_0118

Inhalt

Einleitung

1. Übersicht 1
2. Einzelschrittverfahren und Überrelaxation 4

Überrelaxation bei schiefsymmetrischen Matrizen

3. Überrelaxation bei nicht symmetrischen Matrizen ... 9
4. Matrizen mit $L \leq o$ und $R \geq o$ 18
5. Symmetrische Überrelaxation 23
6. Der optimale Relaxationsfaktor bei $\mathcal{A}$-Matrizen 29
7. Symmetrische Überrelaxation bei $\mathcal{A}$-Matrizen 39

Anwendungen und Ergänzungen

8. Beispiele und Anwendungen 44
9. Ergänzungen 52

Literaturverzeichnis 57

Register

1. Übersicht.

Iterative Verfahren zur Lösung linearer Gleichungssysteme sind zwar seit langem bekannt, sie haben jedoch ihre heutige große Bedeutung erst mit dem Aufkommen programmgesteuerter Rechenmaschinen erlangt. Hier kommen die Vorzüge dieser Verfahren besonders zum Tragen, z.B. Unempfindlichkeit gegenüber Rundefehlern und Ausnutzung der Rechenvorteile, die sich bei Matrizen mit sehr vielen Nullelementen anbieten. Die verbreitete Anwendung löste ihrerseits in den letzten Jahren eine verstärkte Untersuchung der Konvergenzfragen aus.

Ausgangspunkt unserer Untersuchungen ist das klassische Verfahren von Gauß-Seidel oder Einzelschrittverfahren. Erhält man bei diesem den Näherungsvektor $X^{(m+1)}$ aus dem Vektor $X^{(m)}$ durch Addition einer Korrektur $F^{(m)}$, dann verstehen wir unter Überrelaxation die Vorschrift, anstelle von $F^{(m)}$ eine "Überkorrektur" $(1+d)F^{(m)}$ mit einer reellen Zahl d anzusetzen. Diese Modifikation des Einzelschrittverfahrens - die sich während der Rechnung einfach durchführen läßt - kann bei praktischen Rechnungen die Konvergenz sehr verbessern. Ziel der theoretischen Untersuchungen ist es deshalb, den Einfluß der Größe d auf die Konvergenz zu bestimmen.

Dabei bieten sich zwei Fragestellungen an:

1) Läßt sich durch einen geeigneten "Relaxationsfaktor" (1+d) noch Konvergenz in Fällen erzielen, in denen das klassische Verfahren schon divergiert?

2) Läßt sich die Konvergenz durch einen geeigneten Relaxationsfaktor beschleunigen?

Auf beide Fragen lassen sich erst Antworten geben, wenn man über die Koeffizientenmatrix des Gleichungssystems gewisse Voraussetzungen macht. Dabei ist es vorteilhaft, das Gleichungssystem AX = C in der Form

$$X + BX = C$$

zu schreiben und anzunehmen, daß die Diagonalelemente von A gleich 1, d.h. die Diagonalelemente von B gleich o sind.

Ostrowski [20] untersucht symmetrische Matrizen; Kahan [11] setzt voraus, daß die Elemente von B nichtpositiv sind. In beiden Fällen wird die erste Frage verneint (vgl. Abschnitt 2). Dagegen gibt es bei symmetrischen Matrizen fast immer positive d-Werte, die eine Konvergenzbeschleunigung bewirken. Allerdings ist es bis jetzt nicht gelungen, den günstigsten d-Wert, d.h. den "optimalen Relaxationsfaktor" d_o etwa als Funktion gewisser charakteristischer Größen der Matrix B zu bestimmen. Kahan [11] gibt bei symmetrischen nichtpositiven B Schranken für d_o an. Nur für die spezielle Klasse der "$\mathcal{A}$-Matrizen" hat Young [29] den optimalen Relaxationsfaktor als Funktion des Spektralradius von B bestimmen können (vgl. die Bemerkungen zur Literatur in 9.1).

Wir gehen in dieser Arbeit auf den bisher nicht behandelten Fall ein, daß die Matrix B schiefsymmetrisch ist. Hier läßt sich die erste der oben gestellten Fragen bejahen. Es gibt nämlich zu jeder (schiefsymmetrischen) Matrix B ein Intervall von d-Werten, für die das Überrelaxationsverfahren konvergiert,

während beim Einzelschrittverfahren nur für eine bestimmte Klasse von Matrizen Konvergenz eintritt (vgl. Abschnitt 2).

Dieses "Konvergenzintervall" hängt seinerseits von der Reihenfolge ab, in der die Gleichungen durchgerechnet werden. Wir reden bei natürlicher Reihenfolge vom SOR-Verfahren (Successive Overrelaxation Method), bei umgekehrter Reihenfolge vom B-SOR-Verfahren (Backward-SOR-Method); ein SOR- und ein B-SOR-Schritt - alternativ ausgeführt - ergeben das symmetrische Überrelaxationsverfahren oder S-SOR-Verfahren, das schon von Aitken [1], Sheldon [22] u.a. bei symmetrischen Matrizen untersucht wurde. In den Abschnitten 3 bis 5 bestimmen wir Konvergenzintervalle für diese Verfahren.

Der optimale Relaxationsfaktor ist bei schiefsymmetrischen Matrizen im Gegensatz zu symmetrischen i.a. negativ. Präzise Angaben erhalten wir bei schiefsymmetrischen $\mathcal{A}$-Matrizen (Abschnitt 6 und 7). Ein Vergleich mit den analogen Ergebnissen von Young und Kahan für symmetrische $\mathcal{A}$-Matrizen läßt das unterschiedliche Konvergenzverhalten deutlich werden.

In den letzten Abschnitten erläutern wir die Ergebnisse an zwei Beispielen und weisen auf eine Anwendung bei linearen Fredholmschen Integralgleichungen hin. Ferner erwähnen wir nach ergänzenden Bemerkungen zur Literatur einige mit unseren Untersuchungen zusammenhängende Ergebnisse von Stein [23], Weissinger [28], Ostrowski [20], Albrecht [2] und Kredell [15].

Die eingeführten Begriffe und Abkürzungen sind im Anschluß an das Literaturverzeichnis in einem Register zusammengestellt.

2. Einzelschrittverfahren und Überrelaxation.

Nach der Einführung des Einzelschrittverfahrens erwähnen wir die bekannten Konvergenzsätze von Mises-Geiringer-Reich und Stein-Rosenberg und beweisen zwei Varianten dieser Sätze. Anschließend führen wir das SOR-Verfahren ein und formulieren die Konvergenzsätze von Ostrowski und Kahan.

Wir betrachten das lineare Gleichungssystem

$$AX = X + BX = C \ .$$

Dabei sei die Koeffizientenmatrix A vom Grad n, reell und nicht singulär. Wir nehmen an, daß alle Diagonalelemente von A gleich 1 sind (man kann dies o.B.d.A. voraussetzen; vgl. Engeli [6]). Dann sind alle Diagonalelemente von B gleich o. Die linke, untere Dreiecksmatrix von B bezeichnen wir mit L, die rechte, obere Dreiecksmatrix mit R; es ist dann $B = L+R$. Für A ergibt dies die (fast immer benutzte) Zerlegung

$$A=E+B=E+L+R \ .$$

Ist $X^{(o)}$ ein beliebiger Ausgangsvektor, dann lautet die Iterationsvorschrift für das Einzelschrittverfahren

$$X^{(m+1)} = C - LX^{(m+1)} - RX^{(m)} \ , \quad (m=o,1,\ldots) \ .$$

Führt man den "Änderungsvektor" $F^{(m+1)} = X^{(m+1)} - X^{(m)}$ ein, dann erhalten wir (E sei die Einheitsmatrix):

$$(E+L)F^{(m+1)} = -RF^{(m)} \quad \text{oder}$$

$$F^{(m+1)} = -(E+L)^{-1}RF^{(m)} = T_o F^{(m)} \tag{2.1}$$

$$= T_o^m F^{(1)} , \qquad (m=1,2,..) .$$

Wir nennen das Iterationsverfahren konvergent, wenn für jeden Ausgangsvektor $X^{(o)}$ der Vektor $F^{(m+1)}$ gegen den Nullvektor strebt ($m \to \infty$). Dies ist genau dann der Fall (vgl. Householder [10]), wenn jeder Eigenwert t_i von T_o ($i=1,..,n$) dem Betrage nach kleiner als 1 ist, d.h. wenn der

$$\text{Spektralradius} \quad \bar{t}_o = \max_{i=1,..,n} |t_i| < 1$$

ist. Ist diese Bedingung nicht erfüllt, nennen wir das Verfahren divergent. Wir beachten noch, daß T_o von C unabhängig ist, d.h. daß die rechte Seite C des Gleichungssystems für die Frage der Konvergenz ohne Belang ist.

Die Konvergenzbedingung, die von der Spektralradius $\bar{t}_o$ von T_o ausgeht, ist für die Anwendung praktisch [illegible] weil $\bar{t}_o$ im allgemeinen nicht bekannt ist und [illegible] Berechnung aufwendiger wäre als die Lösung des gegebenen Gleichungssystems. Man versucht deshalb, hinreichende Konvergenzkriterien aufzustellen, die mit [illegible] Bekannt sind das Zeilen- und das Spaltensummenkriterium (vgl. Collatz [4]). Andere Bedingungen setzen einen speziellen Typ der Koeffizientenmatrix voraus; wir werden einige dieser Kriterien im folgenden erwähnen.

2.2 Satz (Mises-Geiringer-Reich [17], [21]):

Sei A symmetrisch. Dann gilt:

Das Einzelschrittverfahren konvergiert genau dann, wenn A positiv definit ist.

Uns interessieren vor allem Matrizen $A = D+B$, bei denen $B = L-L^T$ schiefsymmetrisch ist. Zu A gehört nach 2.1 der Operator $T_0(A) = -(E+L)^{-1}(-L^T)$. Führen wir nun die symmetrische Matrix $\tilde{A} = E+\tilde{B} = E+L+L^T$ ein, dann ist $T_0(\tilde{A}) = -(E+L)^{-1}L^T = -T_0(A)$; der Spektralradius $\rho_0(A)$ ist genau dann kleiner als 1, wenn $\rho_0(\tilde{A}) < 1$ ist. Letzteres ist nach 2.1 genau dann der Fall, wenn $\tilde{A}$ positiv definit ist. Also gilt:

2.3 Satz: Sei $A = D+B$ und sei $B = L-L^T$ schiefsymmetrisch; ferner sei $\tilde{B} = L+L^T$. Dann gilt:

Das Einzelschrittverfahren konvergiert genau dann, wenn die symmetrische Matrix $\tilde{A} = E+\tilde{B}$ positiv definit ist.

Stein-Rosenberg [25] leiteten einen anderen Konvergenzsatz über den Begriff der nichtnegativen Matrizen von Perron-Frobenius her.

2.4 Definition: Es ist $B \left\{ \begin{matrix} \geq 0 \\ \leq 0 \end{matrix} \right\}$, wenn die Elemente $b_{ij} \left\{ \begin{matrix} \geq 0 \\ \leq 0 \end{matrix} \right\}$ sind für $i,j = 1,\ldots,n$.

2.5 Satz (Stein-Rosenberg [24]): Sei $A = E+B$, $B \lneq o$. Dann gilt: Das Einzelschrittverfahren konvergiert genau dann, wenn der Spektralradius $\bar{b}$ von B kleiner als 1 ist.

Geht man nun von der Matrix $A = E+L+R$ mit $L \lneq o$ und $R \gneq o$ aus und setzt man $\tilde{B} = L-R$, dann ist $\tilde{B} \lneq o$. Auf die Matrix $\tilde{A} = E+\tilde{B}$ trifft die Voraussetzung von 2.5 zu; beachtet man noch, daß entsprechend wie bei 2.3 $T_o(\tilde{A}) = -T_o(A)$ ist, dann gilt:

2.6 Satz: Sei $A = E+L+R$, $L \lneq o$, $R \gneq o$, $\tilde{B} = L-R$. Dann gilt: Das Einzelschrittverfahren konvergiert genau dann, wenn $\tilde{b} < 1$ ist.

Schreibt man die Iterationsvorschrift für das Einzelschrittverfahren etwas um,

$$X^{(m+1)} = X^{(m)} + (-LX^{(m+1)} - X^{(m)} - RX^{(m)} + C)$$
$$= X^{(m)} + F^{(m)} ,$$

dann erhält man die Vorschrift für das SOR-Verfahren (Successive Overrelaxation Method) aus dem Ansatz

$$X^{(m+1)} = X^{(m)} + (1+d)F^{(m)} \quad , \; d \text{ reell.}$$

Die Zahl (1+d) wird auch Relaxationsfaktor genannt. Man findet:

$$X^{(m+1)} = -(1+d)LX^{(m+1)} - ((1+d)R + dE)X^{(m)} + (1+d)C.$$

Daraus folgt entsprechend wie bei 2.3 für den Operator T_d des SOR-Verfahrens:

$$T_d = -(E+(1+d)L)^{-1}((1+d)R+dE) \; . \tag{2.7}$$

2.8 Lemma: (Kahan [11]): Ist τ_d der Spektralradius von T_d, so ist $\tau_d \geq |d|$.

Beweis: T_d ist das Produkt der beiden Dreiecksmatrizen $D_1 = -(E+(1+d)L)^{-1}$ und $D_2 = ((1+d)R+dE)$. Die Diagonalelemente von D_1 sind -1, die von D_2 sind gleich d. Also ist $\det(T_d) = d^n$. Andererseits ist die Determinante das Produkt der n Eigenwerte t_i von T_d. Daraus folgt $\tau_d = \max |t_i| \geq |d|$ $(i=1,\ldots,n)$.

Für die Konvergenz des SOR-Verfahrens ist also $|d| < 1$ eine notwendige Bedingung. Weiter gilt:

2.9 Satz (Ostrowski [20]): Ist A symmetrisch, dann gilt:
Das SOR-Verfahren konvergiert genau dann, wenn A positiv definit und $-1 < d < 1$ ist.

2.10 Satz (Kahan [11]): Sei $A = E-B$ mit $B \neq 0$.
Ist $\beta \geq 0$, dann divergiert das Gesamtschrittverfahren für keine Wahl von ω. Ist $\beta < 1$, dann konvergiert das SOR-Verfahren für $-1 < d < (1-\beta)/(1+\beta)$.

Eine zusammenhängende Herleitung der in diesem Abschnitt [illegible] S. 56 ff.. Die Sätze von Ostrowski und Kahan stellen Verallgemeinerungen der Ergebnisse von Mises-Geiringer-Reich und Stein-Rosenberg dar. Wir stellen uns nun in den nächsten beiden Abschnitten die Aufgabe, ähnliche Verallgemeinerungen der

3. Überrelaxation bei nicht symmetrischen Matrizen.

Wir untersuchen in diesem Abschnitt das Konvergenzverhalten des SOR- und des B-SOR-Verfahrens bei schiefsymmetrischen Matrizen. Zunächst greifen wir jedoch die von Stein [23], Weissinger [28] und Ostrowski [20] aufgeworfene Frage auf, wie weit sich die Ergebnisse über symmetrische Matrizen (2.2 und 2.9) auf nicht symmetrische Matrizen übertragen lassen. Die erwähnten Verfasser beweisen Sätze - wir formulieren sie als Ergänzung in 9.2 - über "fast-symmetrische" Matrizen. Wir beweisen einen Satz über Matrizen A, deren symmetrischer Anteil $(A+A^T)/2$ positiv definit ist.

Jeder Eigenwert t des Operators T_d besitzt einen normierten Eigenvektor Y, d.h. einen Eigenvektor mit $\bar{Y}Y = 1$. Geht man nun von der Zerlegung $A = E+L+R$ aus und führt man die folgenden Abkürzungen ein: $c = (1-d)/(1+d)$, $a = \bar{Y}AY$ und $s = \bar{Y}(L-R)Y$, dann gilt:

3.1 Lemma: Jeder Eigenwert t von T_d läßt sich in der Form

$$t = \frac{c-a+s}{c+a+s}$$ darstellen.

Beweis: Es ist $tY = T_dY$ oder nach 2.7

$$tY = -(E+(1+d)L)^{-1}((1+d)R+dE)Y ,$$

$$t(E+(1+d)L)Y = -((1+d)R+dE)Y .$$

---

1) Lemma 3.1 ist - ohne explizite Formulierung - in einem Beweis von Albrecht [2] für den Satz 2.9 enthalten.

Diesen Ausdruck benutzen wir für die folgende Umformung:

$$Y^*\{(1-d)E - (1+d)(A+R-L)\}Y = Y^*\{(1-d)E - (1+d)(E+R+L-L+R)\}\cdot Y$$

$$= 2Y^*\{-dE - (1+d)R\}\cdot Y$$

$$= 2tY^*\{E + (1+d)L\}Y$$

$$= tY^*\{2E + (1+d)(-E+E+L+R-R+L)\}\cdot Y$$

$$= tY^*\{(1-d)E + (1+d)(A+L-R)\}\cdot Y.$$

Verwenden wir Anfang und Ende dieser Kette, dann erhalten wir:

$$t = \frac{Y^*\{(1-d)E-(1+d)A+(1+d)(L-R)\}Y}{Y^*\{(1-d)E+(1+d)A+(1+d)(L-R)\}Y}.$$

Dividiert man Zähler und Nenner des Quotienten durch $(1+d)$, benutzt $Y^*Y = 1$ und führt zur Abkürzung $c = (1-d)/(1+d)$, $a = Y^*AY$, $s = Y^*(L-R)Y$ ein, dann erhält man die Behauptung:

$$t = \frac{c - a + s}{c + a + s}.$$

3.2 Definition: Ist G eine symmetrische Matrix und sind $g_1 \leq g_2 \leq \ldots \leq g_n$ die Eigenwerte von G, so sei $l_G = -g_1$, $r_G = g_n$.

Aus der Definition ergibt sich für den Spektralradius $\bar{g}$ von G: $\bar{g} = \mathrm{Max}(l_G, r_G)$.

3.3 Lemma: Ist ein Diagonalelement der symmetrischen Matrix G gleich o, so ist $r_G \geq o$ und $1_G \geq o$.

Beweis: Ist G nicht positiv definit, so ist der kleinste Eigenwert g_1 von G nicht positiv, also $g_1 \leq o$ und $1_G \geq o$. Wenn G nicht negativ definit ist, dann ist der größte Eigenwert $g_n \geq o$ und $r_G = g_n \geq o$. Wir müssen also nur zeigen, daß G nicht definit sein kann, falls ein Diagonalelement von G gleich o ist. O.B.d.A. können wir $g_{11}=o$ annehmen; dann ist $Y^TGY = o$ für $Y^T = (1,o,..,o)$.

Zerlegt man die Matrix $A = E+R+L = E+B$ und setzt man $\tilde{B} = L-R$, dann kann man B und $\tilde{B}$ in symmetrische und schiefsymmetrische Anteile zerlegen:

$$(3.4) \quad B = G + H\,, \quad G = (B+B^T)/2\,, \quad H = (B-B^T)/2.$$
$$\tilde{B} = \tilde{G} + \tilde{H}\,, \quad \tilde{G} = (\tilde{B}+\tilde{B}^T)/2\,, \quad \tilde{H} = (\tilde{B}-\tilde{B}^T)/2.$$

Für die symmetrischen Matrizen G und $\tilde{G}$ lassen sich die Grössen 1_G und $1_{\tilde{G}}$ einführen; ferner seien $\bar{h}$ bzw. $\tilde{\bar{h}}$ die Spektralradien von H und $\tilde{H}$.

3.5 Satz: Ist A eine reelle Matrix und ist der symmetrische Anteil $(A+A^T)/2$ von A positiv definit, so konvergiert das SOR-Verfahren für

$$-1 < d < d_g = \frac{(1-1_G)(1-1_{\tilde{G}}) - \bar{h}\tilde{\bar{h}}}{(1-1_G)(1+1_{\tilde{G}}) + \bar{h}\tilde{\bar{h}}}$$

Beweis: Wir müssen zeigen, daß der Spektralradius $\bar{t}_d$ des Operators T_d im Intervall $-1 < d < d_g$ kleiner als 1 ist.

Nach Lemma 3.1 läßt sich jeder Eigenwert t von T_d mit zugehörigem Eigenvektor Y darstellen in der Form

$$t = \frac{c - a + s}{c + a + s} = \frac{x}{z} .$$

Dabei ist $c = (1-d)/(1+d)$, $a = \overset{*}{Y}AY = \overset{*}{Y}(E+B)Y = 1+\overset{*}{Y}BY = 1+b$ und $s = \overset{*}{X}(L-R)Y = \overset{*}{Y}\overset{*}{B}Y$. Die im allgemeinen komplexen Zahlen b und s lassen sich nach bekannten Sätzen über den "Wertevorrat einer Matrix" (vergl. Zurmühl; wir benutzen $\overset{*}{Y}Y = 1$) einschließen, falls man von den in 3.4 eingeführten Zerlegungen $B = G+iH$ und $\tilde{B} = \tilde{G}+i\tilde{H}$ ausgeht; Ist $b = b_1+ib_2$ und $s = s_1+is_2$ dann gilt:

$$(3.6) \qquad -l_G \leq b_1 \leq r_G , \qquad -\bar{h} \leq b_2 \leq \bar{h} .$$

$$(3.7) \qquad -l_{\tilde{G}} \leq s_1 \leq r_{\tilde{G}} , \qquad -\tilde{\bar{h}} \leq s_2 \leq \tilde{\bar{h}} .$$

Nach Voraussetzung ist der symmetrische Anteil E+G von A positiv definit. Dann ist der kleinste Eigenwert g_1 von G größer als -1, also $-g_1 = l_G < 1$; zusammen mit Lemma 3.3 erhalten wir

$$(3.8) \qquad 0 \leq l_G < 1 .$$

Wir gehen nun von der anfangs eingeführten Darstellung $t = x/z$ aus und berechnen

$$\bar{z}z - \bar{x}x = (c+1+b_1+s_1)^2+(b_2+s_2)^2-(c-1-b_1+s_1)^2-(-b_2+s_2)^2$$

$$= (c+s_1)^2+2(c+s_1)(1+b_1)+(1+b_1)^2+b_2^2+2b_2s_2+s_2^2$$

$$-(c+s_1)^2+2(c+s_1)(1+b_1)-(1+b_1)^2-b_2^2+2b_2s_2-s_2^2$$

$$= 4((c+s_1)(1+b_1)+b_2s_2) = 4f \ .$$

Können wir nun d und damit $c = (1-d)/(1+d)$ so wählen, daß für alle nach 3.6 und 3.7 möglichen Werte von b_1, b_2, s_1, s_2 der Wert f positiv wird, dann ist - für diesen Wert d - $|x|<|z|$, also $|t|<1$ für jeden Eigenwert t von T_d, d.h. $\bar{t}_d<1$.

$f > o$ ist demnach eine hinreichende Konvergenzbedingung; wir schreiben diese etwas um:

$$f \equiv (c+s_1)(1+b_1)+b_2s_2 > o \quad \text{oder}$$

$$c(1+b_1) > -b_2s_2-s_1(1+b_1) \ .$$

Nach 3.6 und 3.8 ist $1+b_1 \geq 1-l_G > o$. Die Ungleichung der vorigen Zeile ist dann gleichbedeutend mit

$$c > \frac{-b_2s_2}{1+b_1} - s_1 \ .$$

Nach 3.6, 3.7 und 3.8 können wir nun abschätzen

$$\frac{-b_2s_2}{1+b_1} - s_1 \leq \frac{\bar{h}\tilde{h}}{1-l_g} + l_{\tilde{G}} \equiv k \ .$$

Die Bedingung $f > o$ ist sicher erfüllt, wenn $c > k$ ist. Schreiben wir $c = (1-d)/(1+d) = -1 + 2/(1+d)$, dann ist $c > k$ gleichbedeutend mit

$$\frac{2}{1+d} > \frac{\bar{h}\tilde{\bar{h}}}{1-l_G} + l_{\tilde{G}} + 1 \quad (> 0) \ ,$$

$$1+d < \frac{2(1-l_G)}{(1-l_G)(1+l_{\tilde{G}})+\bar{h}\tilde{\bar{h}}} \ , \qquad (d > -1) \ .$$

Daraus errechnet sich leicht

$$-1 < d < d_g = \frac{(1-l_G)(1-l_{\tilde{G}}) - \bar{h}\tilde{\bar{h}}}{(1-l_G)(1+l_{\tilde{G}}) + \bar{h}\tilde{\bar{h}}} \ .$$

(die Einschränkung $d > -1$ ergibt sich auch aus Lemma 2.8). Wählt man also für das SOR-Verfahren einen Wert d aus dem offenen Intervall $-1 < d < d_g$, dann wird $|t| < 1$ für jeden Eigenwert t von T_d. Damit ist Satz 3.5 bewiesen. Er besagt u.a., daß bei Matrizen mit positiv definitem symmetrischem Anteil sich immer ein d finden läßt, so daß das SOR-Verfahren konvergiert (es ist immer $d_g > -1$). In 8.1 werden wir noch ein Beispiel zu Satz 3.5 bringen.

Ist A symmetrisch und positiv definit, dann ist auch $B = A-E$ symmetrisch und $\tilde{B} = L-R = L-L^T$ schiefsymmetrisch. Die Anteile H und $\tilde{G}$ sind O (O: Nullmatrix). Dann ist $\bar{h} = l_{\tilde{G}} = 0$. Es wird $d_g = 1$ in Übereinstimmung mit Satz 2.9.

Ist $A = E+B$ und ist $B = L-L^T$ schiefsymmetrisch, dann ist $\tilde{B} = L+L^T$ symmetrisch. Die Anteile G und $\tilde{H}$ sind O, also $H = B$ und $\tilde{G} = \tilde{B}$. Es ist $l_{\tilde{G}} = l_{\tilde{B}}$ und $l_G = \tilde{h} = 0$. Damit folgt aus Satz 3.5 der folgende Satz, der die gesuchte Erweiterung des Satzes 2.3 darstellt:

3.11 Satz: Sei $A = E+B$ und sei $B = L-L^T$ schiefsymmetrisch. Setzt man $\tilde{B} = L+L^T$, dann konvergiert das SOR-Verfahren für

$$-1 < d < d_g = \frac{1-l_{\tilde{B}}}{1+l_{\tilde{B}}} .$$

In dem Ausdruck für die Schranke d_g in Satz 3.5 treten die Größen l_G und $l_{\tilde{G}}$ auf, dagegen nicht r_G und $r_{\tilde{G}}$. Diese etwas willkürlich erscheinende Bevorzugung erklärt sich für l_G, weil $l_G < 1$ gleichbedeutend mit der Voraussetzung ist, der symmetrische Anteil von A sei positiv definit. Dagegen ist das Auftreten von $l_{\tilde{G}}$ durch das spezielle SOR-Verfahren bedingt; durch eine Änderung des Verfahrens tritt $r_{\tilde{G}}$ an die Stelle von $l_{\tilde{G}}$.

Beim SOR-Verfahren haben wir angenommen, daß die Gleichungen in natürlicher Reihenfolge durchgerechnet werden. Entsprechend führen wir nun das zyklische Überrelaxationsverfahren, bei dem die Gleichungen in der umgekehrten Reihenfolge n, n-1,..,1 durchgerechnet werden, als B-SOR-Verfahren ein (Backward-SOR-method; vgl. Kahan [11]).

Beim B-SOR-Verfahren wird die Dreiecksmatrix L mit $X^{(m)}$, die Dreiecksmatrix R mit $X^{(m+1)}$ multipliziert, d.h. die Matrizen L und R treten - verglichen mit dem SOR-Verfahren - mit vertauschten Rollen auf.

Der Ansatz für das B-SOR-Verfahren lautet dann entsprechend der Herleitung für das SOR-Verfahren:

$$(E+(1+d)R)X^{(m+1)} = -((1+d)L+dE)X^{(m)} - (1+d)X^{(m)} + (1+d)C.$$

Der zugehörige Operator ${}_bT_d$ hat die Form

$${}_bT_d = -(E+(1+d)R)^{-1}((1+d)L+dE).$$

Lemma 3.1 gilt in derselben Formulierung auch für ${}_bT_d$; lediglich tritt die Matrix $-B = R-L$ an die Stelle von $B = L-R$.

Hat nun eine symmetrische Matrix G die Eigenwerte $g_1 \leq g_2 \leq \ldots \leq g_n$, dann hat -G die Eigenwerte $-g_n \leq -g_{n-1} \leq \ldots \leq -g_1$. Nach Definition ist

$$l_{-G} = -(-g_n) = g_n = r_G \text{ und } r_{-G} = -g_1 = l_G \text{ oder}$$

$$l_{-G} = r_G , \quad r_{-G} = l_G . \tag{3.13}$$

Für das B-SOR-Verfahren lassen sich nun ähnliche Sätze aufstellen, falls man 3.13 auf die in den Sätzen 3.5 und 3.11 eingeführten Matrizen B und G anwendet.

3.14 Korollar zu 3.5: Ist A eine reelle Matrix und ist der symmetrische Anteil $(A+A^T)/2$ von A positiv definit, dann konvergiert das B-SOR-Verfahren für

$$-1 < d < d_g = \frac{(1-l_G)(1-r_{\tilde{G}}) - h\tilde{h}}{(1-l_G)(1+r_{\tilde{G}}) + h\tilde{h}} .$$

3.15 Korollar zu 3.11: Sei $A = E+B$ und sei $B = L-L^T$ schiefsymmetrisch. Setzt man $\tilde{B} = L+L^T$, so konvergiert das B-SOR-Verfahren für

$$-1 < d < d_g = \frac{(1-r_B)}{(1+r_B)} .$$

Daß sich SOR- und B-SOR-Verfahren bei der gleichen Matrix sehr verschieden verhalten können, werden wir in dem Beispiel 8.2 sehen.

4. Matrizen mit $L \leq o$ und $R \geq o$.

Wir haben im letzten Abschnitt unter bestimmten Voraussetzungen über die Matrix A d-Intervalle $-1 < d < d_g \leq 1$ angegeben, für die das SOR-Verfahren konvergiert und wollen nun die Frage stellen, ob diese Schranken scharf sind, d.h. ob das SOR-Verfahren für $d \geq d_g$ divergiert.

Alle Konvergenzsätze, die ähnlich wie die folgenden Sätze Voraussetzungen über die Vorzeichen der Matrixelemente machen, stützen sich auf einen Satz von Perron-Frobenius aus der Theorie der nichtnegativen Matrizen:

4.1 Satz (Perron-Frobenius): Ist die Matrix $K \geq o$, so existiert ein Eigenwert $k \geq o$ mit $k = \bar{k}$ (Spektralradius von K). Ist $K' \geq K \geq o$, dann ist $\bar{k}' \geq \bar{k} \geq o$.

Setzt man voraus, daß K "unzerlegbar" ist, dann lassen sich wesentlich stärkere Aussagen machen. Ein Beweis mit Literaturhinweisen auf andere Beweise findet sich bei Varga [26] (p. 25 ff.).

4.2 Lemma: τ ist genau dann Eigenwert des Operators T_d, wenn $(t+d)/(1+d)$ Eigenwert der Matrix $B_t = -(tL+R)$ ist.

Beweis: Die Eigenwerte t_j von T_d ($j=1,\ldots,n$) sind die Nullstellen des charakteristischen Polynoms von T_d, d.h. die Nullstellen der Determinante

$$\det\{tE - (E+(1+d)L)^{-1}(-(1+d)R-dE)\}$$

$$= \det(E+(1+d)L)^{-1}\det\{t(E+(1+d)L)+(1+d)R+dE\}$$

$$= (1+d)^n\det\left\{\left(\frac{t+d}{1+d}\right)E + (tL+R)\right\} \quad .$$

Am Schluß der Kette steht das charakteristische Polynom der Matrix $B_t = -(tL+R)$. Ist nun $t = t_j$ eine Nullstelle des ersten Polynoms, d.h. Eigenwert von T_d, dann ist $b_j = (t_j+d)/(1+d)$ eine Nullstelle des letzten Polynoms, d.h. Eigenwert von B_t. Derselbe Schluß gilt auch in der umgekehrten Richtung.

4.3 Satz: Sei $A = E+L+R$ und sei $L \leq o$ und $R \geq o$. Setzt man $\tilde{B} = L-R$ und ist $\tilde{\bar{b}} < 1$, dann divergiert das SOR-Verfahren für

$$d \geq \frac{1-\tilde{\bar{b}}}{1+\tilde{\bar{b}}} \quad .$$

Beweis: Wir müssen zeigen, daß für $d \geq (1-\tilde{\bar{b}})/(1+\tilde{\bar{b}})$ der Spektralradius $\bar{t}_d$ von T_d größer als 1 ist.
Nach Voraussetzung ist $\tilde{B} \leq o$. Es existiert nach 4.1 ein reeller Eigenwert $-\tilde{\bar{b}} \leq o$.
Setzen wir in Lemma 4.2 $t = -1$, dann erhalten wir: $t = -1$ ist ein Eigenwert von T_d, wenn $(-1+d)/(1+d)$ Eigenwert von $B_{-1} = -(-L+R) = L-R = \tilde{B}$ ist. Wir wählen nun $d = d'$ so, daß $-\tilde{\bar{b}} = (-1+d')/(1+d')$ wird und lösen nach d' auf: $d' = (1-\tilde{\bar{b}})/(1+\tilde{\bar{b}})$
Für diesen Wert d' hat T_d einen Eigenwert $t = -1$; es ist $\bar{t}_{d'} \geq 1$.

Wir müssen nun noch zeigen, daß für alle $d \geq d$ auch $\bar{t}_d \geq \bar{t}_{d'} \geq 1$ ist. Setzt man $(1+d)L = -K$, dann ist nach Voraussetzung $K \geq o$ für $|d| < 1$. Es ist ferner

$$(E+(1+d)L)^{-1} = (E-K)^{-1} = E+K+K^2+\ldots+K^{n-1} \geq o \quad .$$

Daraus folgt

$$(4.4) \qquad T_d = -(E+(1+d)L)^{-1}((1+d)R+dE) \leq o$$

[illegible] $L \leq o$, $R \geq o$, $d \geq o$. Nach [illegible]

$$(4.5) \qquad T_{d'} \leq T_d \leq [illegible] \text{ für } d \geq d' \geq o \quad .$$

Nach Voraussetzung ist nun $\tilde{\bar{b}} \leq 1$; also wird der oben eingeführte Wert $d' = (1-\tilde{\bar{b}})/(1+\tilde{\bar{b}}) \geq o$. Nach 4.5 ist $\bar{t}_d \geq \bar{t}_{d'} \geq 1$ für $d \geq d'$. Damit ist Satz 4.3 bewiesen.

4.6 Korollar: Sei $A = E+L-L^T$ und sei $L \leq o$. Setzt man $\tilde{B} = L+L^T$ und ist $\tilde{\bar{b}} \leq 1$, dann konvergiert das SOR-Verfahren für $-1 < d < d_g = (1-\tilde{\bar{b}})/(1+\tilde{\bar{b}})$ und divergiert für alle anderen Werte von d.

Beweis: Nach Definition von $l_{\tilde{B}}$ und nach 4.1 ist für $\tilde{B} \leq o$ $\tilde{\bar{b}} = l_{\tilde{B}}$. Damit folgt die erste Behauptung aus Satz 3.11, die zweite aus 4.3.

Korollar 4.6 gilt vermutlich auch für $\tilde{b} > 1$; in diesem Fall können wir uns jedoch nicht auf 4.5 stützen, so daß wir nur behaupten können:

(4.7) Das SOR-Verfahren divergiert für $d = (1-\tilde{b})/(1+\tilde{b})$.

Ist $L \geq 0$ und $R \leq 0$ - sind also die Verhältnisse bezüglich L und R vertauscht - , dann lassen sich 4.5 und 4.6 für das B-SOR-Verfahren aussprechen.

Nachdem sich auf diese Weise für gewisse Matrizen ein d-Intervall angeben läßt, in dem das SOR-Verfahren konvergiert, erhebt sich die Frage, für welche d-Werte sich eine "schnelle" Konvergenz erzielen läßt, d.h. zur Erzielung einer bestimmten Genauigkeit möglichst wenig Iterationen notwendig sind. Dies ist dann der Fall, wenn der Spektralradius $\bar{\tau}_d$ von T_d möglichst klein wird (vgl. z.B. Engeli [5]).

Nun ist $\bar{\tau}_d$ bei gegebener Matrix A eine stetige Funktion von d. Ferner wissen wir nach Lemma 2.8, daß für $|d| \geq 1$ auch $\bar{\tau}_d \geq 1$ ist, während es im Bereich $|d| < 1$ Werte von d gibt, für die $\bar{\tau}_d < 1$ ist (solange das Verfahren konvergiert, und nur dieser Fall ist für uns von Interesse).

Damit können wir den optimalen Relaxationsfaktor d_o einführen durch die Forderung:

$$\bar{\tau}_{d_o} = \min_d (\bar{\tau}_d) \tag{4.8}$$

(wird das Minimum für mehrere Werte von d angenommen, dann sei d_o der dem Betrage nach kleinste Wert).

4.9 Satz: Ist $A = E+L+R$, $L \leq o$, $R \geq o$, dann ist $d_o \leq o$.

Beweis: Ist $\tilde{B} = L-R$ und ist $\tilde{b} > 1$, dann ist $d_g < o$ nach 4.7 und damit auch $d_o < o$. Ist $\tilde{b} \leq 1$, dann ist nach 4.5: $T_d \leq T_{d'} \leq o$ und $\bar{\tau}_d \geq \bar{\tau}_{d'}$ für $d \geq d' \geq o$. Dies ist die Behauptung.

Numerische Beispiele lassen vermuten, daß bei schiefsymmetrischen Matrizen immer $d_o \leq o$ ist, während bei symmetrischen Matrizen i.a. $d_o \geq o$ ist (allerdings zeigt Kahan [11], daß es symmetrische Matrizen gibt, für die $d_o < o$ ist). Wir werden in den Abschnitten 6 und 7 noch ausführlicher auf die Grösse d_o eingehen.

5. Symmetrische Überrelaxation.

Beim symmetrischen Überrelaxationsverfahren oder kurz S-SOR-Verfahren werden ein SOR- und ein B-SOR-Schritt in alternierender Reihenfolge ausgeführt. Eine Iteration des S-SOR-Verfahrens besteht also darin, die Gleichungen zuerst in der Reihenfolge 1,2, ..., n und anschließend in der Reihenfolge n, n-1,..., 1 durchzurechnen. Eine Untersuchung des Konvergenzverhaltens ist erforderlich, weil die Konvergenz des SOR- und des B-SOR-Verfahrens weder notwendig noch hinreichend für die Konvergenz des S-SOR-Verfahrens ist (vgl. Abschnitt 7).

Für d = o ist das Verfahren schon von Aitken [1] als Erweiterung des Einzelschrittverfahrens beschrieben worden. Das allgemeinere S-SOR-Verfahren wurde von Sheldon [22], Lanan [10] und Habetler-Wachspress [9] bei symmetrischen Matrizen untersucht. Wir werden in diesem Abschnitt im wesentlichen einen Konvergenzsatz für schiefsymmetrische Matrizen beweisen.

Ist AX = C das gegebene Gleichungssystem und ist A = E+L+R, dann erhält man beim SOR-Verfahren für den Änderungsvektor

$$F^{(m+1)} = X^{(m+1)} - X^{(m)} :$$

$$F^{(m+1)} = T_d F^{(m)} = -(E+(1+d)L)^{-1}((1+d)R+dE)F^{(m)}, \quad (m=o,1..).$$

und entsprechend beim B-SOR-Verfahren

$$F^{(m+2)} = {}_bT_dF^{(m+1)} = -(E+(1+d)R)^{-1}((1+d)L+dE)F^{(m+1)} .$$

Beim S-SOR-Verfahren ergeben dann beide Schritte zusammen:

$$(5.2)\quad F^{(m+2)} = (E+(1+d)R)^{-1}((1+d)L+dE)(E+(1+d)L)^{-1}((1+d)R+dE)F^{(m)}$$

$$= T_{dd}F^{(m)}, \qquad (m = 0,2,4,\ldots).$$

T_{dd} ist der Operator des S-SOR-Verfahrens. T_{dd} besteht aus 4 Faktoren; dabei sind der zweite und dritte miteinander vertauschbar. Die Determinante jedes Faktors ist gleich dem Produkt der Diagonalelemente. Es ist also

$$\det(T_{dd}) = d^{2n}.$$

Andererseits ist das Produkt der n Eigenwerte von T_{dd} gleich $\det(T_{dd})$. Es folgt:

5.2 Lemma: Für den Spektralradius $\bar{t}_{dd}$ des Operators gilt:

$$\bar{t}_{dd} \geq d^2.$$

Da das S-SOR-Verfahren genau dann konvergent ist, wenn $\bar{t}_{dd} < 1$ ist, können wir nach 5.3 weiterhin $|d|<1$ voraussetzen.

Wir erwähnen zunächst ein bekanntes Ergebnis:

5.4 Satz (Ostrowski [20], Kahan [11], Habetler-Wachspress [9]):
Ist A symmetrisch, dann gilt:
Das S-SOR-Verfahren konvergiert genau dann, wenn
A positiv definit und $-1 < d < 1$ ist.

Beim Beweis gehen Kahan und Habetler-Wachspress von T_{dd} aus, während Ostrowski einen Satz über allgemeine Überrelaxation beweist, der die Sätze 2.9 und 5.4 einschließt (er läßt eine freie Auswahl in der Reihenfolge der Gleichungen zu).

Jeder Eigenwert t von T_{dd} besitzt einen normierten Eigenvektor Y mit $\bar{Y}Y = 1$. Geht man nun von der Zerlegung $A = E+L+R$ aus und führt man die Abkürzungen $\sigma = \bar{Y}LRY$ und $\beta = \bar{Y}(L+R)Y$ ein, dann gilt:

5.5 Lemma: Jeder Eigenwert t von T_{dd} läßt sich in der Form

$$t = \frac{d^2 + (1+d)^2\sigma + d(1+d)\beta}{1 + (1+d)^2\sigma + (1+d)\beta}$$ darstellen.

Beweis: Wir gehen von der Darstellung 5.2 aus und vertauschen den zweiten und dritten Faktor:

$$\begin{aligned} T_{dd} &= (E+(1+d)R)^{-1}(E+(1+d)L)^{-1}((1+d)L+dE)((1+d)R+dE) \\ &= \{(E+(1+d)L)(E+(1+d)R)\}^{-1}\{(1+d)^2LR+d(1+d)(L+R)+d^2E\} \\ &= \{E+(1+d)(L+R)+(1+d)^2LR\}^{-1}\{(1+d)^2LR+d(1+d)(L+R)+d^2E\}. \end{aligned}$$

Ist t ein Eigenwert von T_{dd} und Y ein zugehöriger Eigenvektor mit $\bar{Y}Y = 1$, dann folgt aus $T_{dd}Y = tY$:

$$\bar{Y}\{(1+d)LR+d(1+d)(L+R)+d^2E\}Y = t\bar{Y}\{E+(1+d)(L+R)+(1+d)^2LR\}Y$$

oder
$$t = \frac{\bar{Y}\{(1+d)^2LR+d(1+d)(L+R)+d^2E\}Y}{\bar{Y}\{(1+d)^2LR+(1+d)(L+R)+E\}Y} .$$

Multipliziert man die Klammern im Zähler und Nenner aus und setzt man $\bar{Y}LRY = \sigma$, $\bar{Y}(L+R)Y = \beta$, dann folgt die Behauptung.

1.5 Satz: Sei $A = E+B$ und sei $B = L+L^T$ [illegible].
Setzt man $\tilde{B} = L+L^T$ und $c = b^2+\tilde{b}^2$, dann gilt:

a) Ist $c \leq 1$, dann konvergiert das S-SOR-Verfahren für [illegible]

b) Ist [illegible], dann [illegible] für

[illegible]

Beweis: [illegible] $R = -L^T$ und $\sigma = -\bar{Y}LL^TY = -s$ mit $s \geq 0$, ferner $\beta = \bar{Y}(L-L^T)Y = ib$ (b reell). Jeder Eigenwert t von T_{dd} läßt sich also in der Form

$$t = \frac{d^2-(1+d)^2s+(1+d)db\cdot i}{1-(1+d)^2s+(1+d)b\cdot i} = \frac{x_1+x_2i}{[illegible]} \qquad (s > 0,\ b \text{ reell})$$

darstellen. Es ist stets $|x_2| < |y_2|$ für $|d|<1$, also ist $|t|<1$, falls auch noch $|x_1| = |d^2-(1+d)^2 s| < |y_1| = |1-(1+d)^2 s|$ wird. Es genügt zu zeigen: $y_1+x_1 > o$ und $y_1-x_1 > o$.
Es ist

$$(5.7) \qquad y_1-x_1 = 1-d^2 > o \quad \text{für } |d|<1.$$

Um die zweite Beziehung abzuleiten, schätzen wir zunächst die Größe s ab. Setzen wir $\overset{*}{Y}LL^T Y = s \geq o$, $\overset{*}{Y}L^T LY = r \geq o$ und $\overset{*}{Y}(L^{<}+L^{T^2})Y = v$, dann gilt (Wir benutzen die Sätze über den Wertebereich einer Matrix; vergl. 3.7 und 3.8):

$$o \leq \overset{*}{Y}BB^T Y = -\overset{*}{Y}B^2 Y = -\overset{*}{Y}(L-L^T)(L-L^T)Y = s+r-v \leq \bar{b}^2.$$

$$o \leq \overset{*}{Y}\tilde{B}\tilde{B}^T Y = \overset{*}{Y}\tilde{B}^2 Y = \overset{*}{Y}(L+L^T)(L+L^T)Y = s+r+v \leq \tilde{\bar{b}}^2.$$

Beide Gleichungen zusammen ergeben $2s+2r \leq \bar{b}^2+\tilde{\bar{b}}^2$ oder, da $r \geq o$ ist,

$$(5.8) \qquad o \leq 2s \leq \bar{b}^2+\tilde{\bar{b}}^2 = c \ .$$

Dann ist

$$(5.9) \qquad y_1+x_1 = 1+d^2-2s(1+d)^2 \geq 1+d^2-c(1+d)^2.$$

Ist $c \leq 1$, dann wird $y_1+x_1 \geq 1+d^2-(1+d)^2 = -2d > o$ für $d < o$. Also wird $|t|<1$ für $-1 < d < o$; damit ist Behauptung a) bewiesen.

Weiterhin sei $c > 1$. Dann bestimmen wir die Nullstellen der in d quadratischen Gleichung

$$q(d) \equiv (1+d^2) - c(1+d)^2 = 0 ,$$

$$d^2(c-1) + 2cd + (c-1) = 0 ,$$

$$d_{1,2} = \frac{-c \pm \sqrt{c^2 - c^2 + 2c - 1}}{c-1}$$

$$= -1 \pm \frac{\sqrt{2c-1}}{c-1} \quad \mp 1 .$$

Für alle d-Werte innerhalb dieser beiden Nullstellen wird $q(d) > 0$, also auch nach 5.9 $y_1 + x_1 \geq q(d) > 0$. Zusammen mit 5.7 erhalten wir: Ist $c > 1$, dann wird $|t| < 1$ für

$$-1 < d < -1 + \frac{\sqrt{2c-1} \quad -1}{c-1} .$$

Damit ist die Behauptung b) bewiesen.

Die Größe $(\sqrt{2c-1} \quad -1)/(c-1)$ strebt gegen 0 für $c \to \infty$, d.h. das "Konvergenzintervall" wird für größer werdendes c immer kleiner.

Die gleiche Größe strebt gegen 1 für $c \to 1$, wie sich durch Anwenden der Regel von de l'Hôpital erkennen läßt, d.h. das Konvergenzintervall geht für $c \to 1$ in den Bereich $-1 < d < 0$ über (in Übereinstimmung mit der ersten Behauptung von 5.6).

6. Der optimale Relaxationsfaktor bei $\mathcal{A}$-Matrizen.

Über den optimalen Relaxationsfaktor d_o (vgl. 4.8) ist selbst bei symmetrischen Matrizen nicht viel bekannt. Kahan [11] gibt Abschätzungen an. Young [29] hat für eine spezielle Klasse von Matrizen, die er Matrizen mit "Property A" und "consistent ordered" nannte, im symmetrischen Fall genaue Ergebnisse erhalten. Wir werden im folgenden einen entsprechenden Satz für schiefsymmetrische Matrizen herleiten und das unterschiedliche Verhalten bei symmetrischen und schiefsymmetrischen Matrizen an einem Schaubild erläutern.

6.1 Definition: Die Matrix $A = E+B$ ist eine $\mathcal{A}$-Matrix, wenn B die Gestalt

$$B = \begin{pmatrix} 0_1 & F \\ D & 0_2 \end{pmatrix}$$

hat (0_1, 0_2 quadratische Nullmatrizen).

Ist A eine $\mathcal{A}$-Matrix und schreibt man das lineare Gleichungssystem $AX = C$ in der Form $X = -BX+C$, dann kann man den Vektor X in zwei Teilvektoren X_1 und X_2 zerlegen (entsprechend auch C in C_1 und C_2), so daß das Gleichungssystem die Gestalt

Weise nur von Komponenten des Vektors X_2 ab und umgekehrt. α-Matrizen können bei der numerischen Lösung von elliptischen partiellen Differentialgleichungen auftreten (häufig müssen noch gewisse Zeilen und entsprechende Spalten vertauscht werden, ehe das Gleichungssystem die Form 6.2 hat).

6.3 Lemma: Ist $A = E+L+R$ eine α-Matrix, dann ist

$$L^2 = R^2 = 0 .$$

Beweis: Man geht von 6.1 aus und bestätigt das Ergebnis durch Ausrechnen.

6.4 Lemma: Ist $A = E+R+L$ eine α-Matrix, so ist

$$\det(xE-yL-zR) \equiv \det(xE \pm \sqrt{yz}\,L \pm \sqrt{yz}\,R)$$

für beliebige Zahlen $x \neq 0$, y, z.

Beweis: Setzt man $\alpha = (y \pm \sqrt{yz})/x$ und $\beta = (z \pm \sqrt{yz})/x$, so ist

$$\det(xE-yL-zR) \equiv \det(E+\alpha L)(xE-yL-zR)(E+\beta R)$$

$$\det(xE \pm \sqrt{yz}\,L \pm \sqrt{yz}\,R) ,$$

wie sich durch Ausrechnen des zweiten Ausdrucks und Benutzung von 6.3 bestätigen läßt.

6.5 Korollar: Wenn $A = E+B$ eine $\mathfrak{A}$-Matrix ist, dann kommen die Eigenwerte von B immer paarweise in der Form $\pm b_j$ $(j=1,..,m)$ vor. Ist $B = L+L^T$ symmetrisch und sind $\pm b_j$ die reellen Eigenwerte von B, so sind $\pm ib_j$ $(j=1,..,m)$ die rein imaginären Eigenwerte der schiefsymmetrischen Matrix $B = L-L^T$.

Beweis: Für $y = z = 1$ steht in 6.4 links die charakteristische Determinante von B und rechts für das untere Vorzeichen die charakteristische Determinante von -B. Da nach 6.4 beide identisch sind, haben B und -B dieselben Eigenwerte. Dies ist gleichbedeutend mit der ersten Behauptung. Die zweite Behauptung folgt aus 6.4 für $y = 1$, $z = -1$.

6.6 Satz (Young [29]): Ist $A = E+B$ eine $\mathfrak{A}$-Matrix, so entspricht jedem Paar $\pm b_j$ von Eigenwerten der Matrix B ein Paar von Eigenwerten $t_1^{(j)}$, $t_2^{(j)}$ des Operators T_d, das der quadratischen Gleichung

$$(t+d)^2 - t(1+d)^2 b_j^2 = 0 \ , \ (j=1,..,m),$$

genügt.

Beweis: Die Eigenwerte t_j des Operators T_d sind die Nullstellen der Determinante [1)]

---

1) Der von uns vorgeführte Beweis, der Lemma 6.4 benutzt, weicht von anderen Darstellungen ab und geht auf eine Beweisskizze von Kahan [11] zurück.

$$\det(tE-T_d) \equiv \det\{tE + (E+(1+d)L)^{-1}((1+d)R+dE)\}$$

$$\equiv \det(E+(1+d)L)^{-1}\det(t(E+(1+d)L)+(1+d)R+dE)$$

$$\equiv \det\{(t+d)E + t(1+d)L + (1+d)R\} \quad .$$

Nach Lemma 6.4 gilt:

$$\det\{(t+d)E+t(1+d)L+(1+d)R\} \equiv \det\{(t+d)E\pm(1+d)\sqrt{t}L\pm(1+d)\sqrt{t}R\}$$

$$\equiv \det\{(t+d)E \pm (1+d)\sqrt{t}B\} \; .$$

Ist nun $\pm b_j$ ein Eigenwertpaar von B, dann gilt nach der letzten Beziehung

$$(t+d) = \pm(1+d)\sqrt{t}b_j \qquad \text{oder}$$

$$(t+d)^2 - t(1+d)^2 b_j^2 = 0 \quad .$$

Entsprechend der Herleitung sind die beiden Nullstellen $t_1^{(j)}$, $t_2^{(j)}$ dieser quadratischen Gleichung Eigenwerte von T_d.

Für das Einzelschrittverfahren gilt nach 6.6: Zu jedem Eigenwertpaar $\pm b_j$ von B gehörten die Eigenwerte $t_1^{(j)} = 0$ und $t_2^{(j)} = b_j^2$ von T_0. Es ist $\bar{t}_0 = \bar{b}^2$. Beachtet man, daß $\bar{b}$ der Spektralradius des Operators des Gesamtschrittverfahrens ist, und wählt man $-\ln(\bar{t}_0)$ als Maß für die Konvergenzgeschwindigkeit (vergl. Young [29]), dann läßt sich dieses Ergebnis auch so aussprechen: Bei $\mathcal{A}$-Matrizen konvergiert das Einzelschrittverfahren doppelt so schnell wie das Gesamtschritt-

6.7 Satz (Young [29]): Ist $A = E+B$ eine symmetrische, positiv definite $\mathfrak{A}$-Matrix, dann gilt:

a) Das SOR-Verfahren konvergiert für $-1 < d < 1$.

b) Der optimale Relaxationsfaktor ist $d_o = \left(\frac{\bar{b}}{1 + \sqrt{1-\bar{b}^2}}\right)^2$.

c) Für $d \neq d_o$ ist der Spektralradius $\bar{\tau}_d > \bar{\tau}_{d_o}$.

d) Für $d \geq d_o$ ist $\bar{\tau}_d = d$.

Beweise finden sich bei Young und z.B. Varga [26](S. 105 ff.). Teil a) folgt aus 2.9. Ein Beweis von 6.7 ließe sich auch nach der Methode durchführen, die wir für den Beweis des folgenden Satzes anwenden.

6.8 Satz: Ist $A = E+B$ eine $\mathfrak{A}$-Matrix und ist B schiefsymmetrisch, dann gilt:

a) Das SOR-Verfahren konvergiert für $-1 < d < \frac{1-\bar{b}}{1+\bar{b}}$.

b) Der optimale Relaxationsfaktor ist $d_o = -\left(\frac{\bar{b}}{1 + \sqrt{1+\bar{b}^2}}\right)^2$.

c) Für $d \neq d_o$ ist der Spektralradius $\bar{\tau}_d > \bar{\tau}_{d_o}$.

d) Für $d \leq d_o$ ist $\bar{\tau}_d = |d|$.

Beweis: Die Eigenwerte der schiefsymmetrischen Matrix B sind alle rein imaginär, also in der Form $\pm ib_j$ (b_j reell; $j = 1,..,m$) darstellbar. Fassen wir zunächst $b_j = b$ als Variable auf - wir können o.B.d.A. $b \geq 0$ voraussetzen - dann

erhalten wir aus 6.6

$$(t+d)^2 + t(1+d)^2b^2 = o \ .$$

Wir lösen nach t auf:

$$2t_{1,2} = -2d-(1+d)^2b^2 \pm \sqrt{(2d+(1+d)^2b^2)^2 - 4d^2} \qquad \text{oder}$$

$$(6.9) \quad 2t_{1,2} = -2d-(1+d)^2b^2 \pm (1+d)b\sqrt{(1+d)^2b^2 + 4d} \ .$$

Für $(1+d)^2b^2 + 4d \equiv f < o$ werden $t_{1,2}$ konjugiert komplex. Man rechnet leicht nach:

$$(6.10) \qquad |t_{1,2}| = |d| \text{ für } f \leq o \ .$$

Weiterhin sei $f > o$. Dann ist in 6.9

$$(6.11) \qquad -(2d+(1+d)^2b^2) = -f+2d < o \ ,$$

wie man für $d \geq o$ aus dem ersten Teil und für $d < o$ aus dem zweiten Teil der Beziehung 6.11 erkennt. In 6.9 ist demnach $t_2 < o$ und $|t_2| > |t_1|$. Da uns nur die betragsgrößte Wurzel interessiert, setzen wir $t_2 = t$ und erhalten

$$2t = -((1+d)^2b^2+2d)-(1+d)bf^{1/2} \ .$$

Es ist $\frac{\partial t}{\partial b} = -(1+d)^2b - (1+d)(f-2d)f^{-1/2} < o$ nach 6.11.

Da t selbst negativ ist, folgt: Bei konstantem d wächst $|t|$, falls b wächst. Der größte Wert, den t in unserem Fall annehmen kann, entspricht dem Spektralradius $\bar{\rho}$ von L. Dieses Ergebnis läßt sich auch so aussprechen:

(6.12) Der betragsgrößte, reelle Eigenwert von T_d ist durch $\bar{b}$ bestimmt.

Ferner ist $\frac{\partial t}{\partial d} = -\{1+(1+d)b+b(f+1-d)f^{-1/2}\} < 0$ für $|d|<1$, also gilt (t ist negativ):

(6.13) t wächst monoton, falls d wächst ($f>0$).

Wir kehren nun zu unserem Ausgangsproblem zurück, bei gegebener Matrix B und damit gegebenem $\bar{b}$ den Wert d_0 zu bestimmen. 5.11, 5.12 und 5.13 zusammen ergeben: $|t|$ wird minimal, wenn für $b = \bar{b}$ der Wert $f = 0$ ist. Also gilt:

$$(1+d_0)^2\bar{b}^2 + 4d_0 = 0 .$$

Eine der beiden Lösungen ist kleiner als -1; es kommt nur die andere in Betracht:

$$d_0 = \frac{-(\bar{b}^2+2)+2\sqrt{\bar{b}^2+1}}{\bar{b}^2}$$

$$= -\frac{(\bar{b}^2+1) - 2\sqrt{\bar{b}^2+1} + 1}{\bar{b}^2}$$

$$= -\left(\frac{\bar{b}}{1+\sqrt{1+\bar{b}^2}}\right)^2 .$$

Damit ist die Behauptung 6.8b) bewiesen. Aus 6.10 und 6.11 folgen die Behauptungen c) und d).

Um 6.8a zu beweisen, bestimmen wir die Werte von d, für die (bei gegebenem $\bar{b}$) $|t|\geq 1$ wird. Nach 6.10 kann - für $|d|<1$ -

nur für reelle Wurzeln von 6.9 $|t| \geq 1$ werden; es sei $t = -1$ für $d = d_g$ und $b = \bar{b}$. Dies ergibt die Gleichung:

$$-2 = -2d_g - (1+d_g)^2\bar{b}^2 - (1+d_g)\bar{b}\sqrt{(1+d_g)^2\bar{b}^2 + 4d_g} \quad .$$

Es liegt nur eine Lösung in dem Intervall $-1 < d_g < 1$:

$$d_g = \frac{1-\bar{b}}{1+\bar{b}} \quad .$$

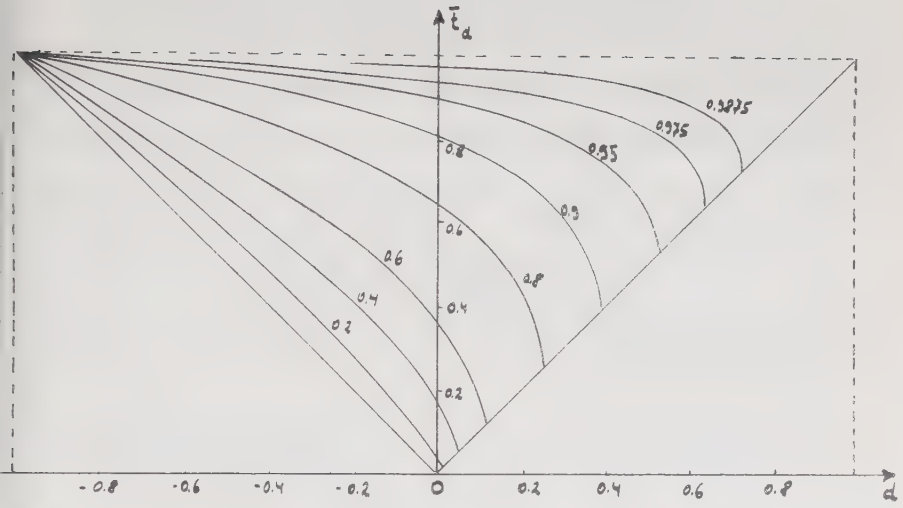
Aus 6.13 folgt: Das SOR-Verfahren konvergiert für $-1 < d < d_g$ und divergiert für $d \geq d_g$. Nach Satz 3.11 erhält man dasselbe Ergebnis, wenn man benutzt, daß nach 6.5 bei $\mathfrak{A}$-Matrizen $l_B = r_B = \bar{b}$ ist.

Bemerkung: Beim B-SOR-Verfahren erhält man die Eigenwerte des Operators ${}_bT_d$ als Nullstellen der Determinante

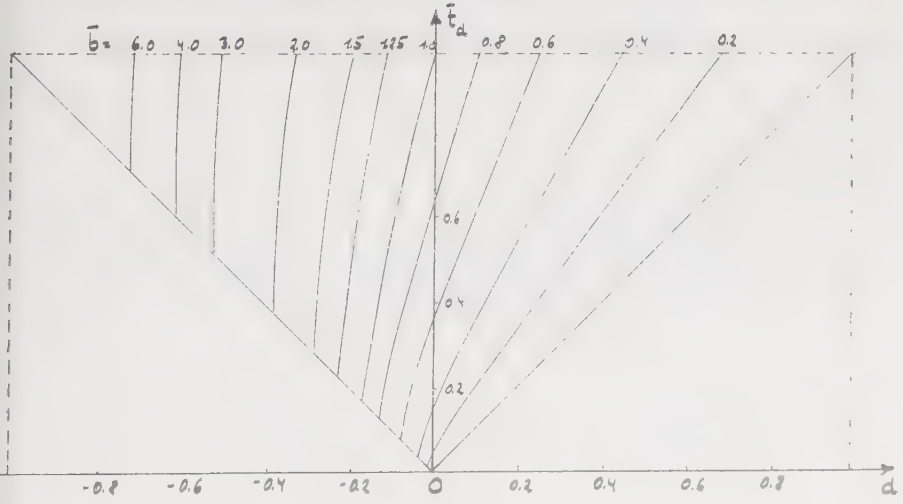
$$\det(tE - {}_bT_d) \equiv \det\{(t+d)E + (1+d)L + t(1+d)R\} \quad .$$

Bei $\mathfrak{A}$-Matrizen ist diese Determinante nach Lemma 6.4 identisch mit $\det(tE - T_d)$. SOR- und B-SOR-Verfahren verhalten sich also bei $\mathfrak{A}$-Matrizen - was die Konvergenz anbelangt - völlig gleich.

Um das unterschiedliche Konvergenzverhalten bei symmetrischen (Satz 6.7) und schiefsymmetrischen (Satz 6.8) Matrizen herauszustellen, sind in Abbildung 1 die Spektralradien $\bar{t}_d$ als Funktion von d für verschiedene Parameterwerte $\bar{b}$ aufgetragen. Im symmetrischen Fall ist das Verfahren nur für $\bar{b} < 1$



a) B symmetrisch



b) B schiefsymmetrisch

Abb. 1: SOR-Verfahren bei $\mathcal{M}$-Matrizen A = E+B

$\bar{t}_d$ als Funktion von d (Parameter $\bar{b}$)

konvergent, allerdings im ganzen Intervall $-1 < d < 1$. Bei schiefsymmetrischen Matrizen läßt sich für alle Werte von $\bar{b}$ Konvergenz erreichen, jedoch nur für $-1 < d < (1-\bar{b})/(1+\bar{b})$; z.B. divergiert das SOR-Verfahren für $\bar{b} = 1$ bei symmetrischen Matrizen; ebenso divergiert das Einzelschrittverfahren ($d = o$) bei schiefsymmetrischen Matrizen, während beim SOR-Verfahren $d_o \approx -o.172 = -\bar{t}_{d_o}$ wird. Das Verfahren konvergiert im Mittel etwa so schnell wie die geometrische Reihe mit dem Faktor o.172.

[illegible] bei praktischen Anwendungen d_o zu bestimmen, ben[illegible] man [illegible]. diese Größe ist selten exakt bekannt. Man benutzt deshalb häufig obere Schranken $\bar{t} \geq \bar{b}$, die ein etwas verfälschtes $|d_o| \geq |d_o|$ ergeben. Die Abbildung zeigt, daß eine Verfälschung in dieser Richtung die Konvergenz nicht so stark beeinflußt wie ein $|d_o| < |d_o|$. Garré [?] und Kulsrud [10] zeigen, wie [illegible] Größen, die während der Rechnung anfallen, Abschätzungen für $\bar{b}$ erhalten lassen. [illegible] [?] weist darauf hin, daß man aus dem Verhalten der [illegible] $\bar{x}^{(m+1)} - \bar{x}^{(m)}$ [illegible] $\bar{x}^{(m)}$ erkennen kann, ob der zum Spektralradius $\bar{t}_1$ gehörige Eigenwert reell oder komplex ist, d.h. ob $|d| \geq |d_o|$ oder $|d| < |d_o|$ ist.

7. Symmetrische Überrelaxation bei $\mathcal{A}$-Matrizen.

Auch beim S-SOR-Verfahren interessieren die Werte $d = d_{oo}$, für die bei einer vorgegebenen Matrix die Konvergenz optimal wird (d_{oo} wird genauso definiert wie d_o in 4.8). Genaue Ergebnisse erhält man auch hier für $\mathcal{A}$-Matrizen. Kahan [11] zeigt, daß bei symmetrischen $\mathcal{A}$-Matrizen $d_{oo}=o$ wird. Wir zeigen, daß bei schiefsymmetrischen Matrizen $d_{oo} \neq o$ ist und vergleichen dann mit dem Ergebnis des letzten Abschnitts.

7.1 Lemma (Kahan): Ist $A = E+B$ eine $\mathcal{A}$-Matrix, dann entspricht jedem Paar $\pm b_j$ $(j=1,..,m)$ von Eigenwerten der Matrix B ein Paar von Eigenwerten $t_1^{(j)}$, $t_2^{(j)}$ von Eigenwerten des Operators T_{dd}. $t_1^{(j)}$ und $t_2^{(j)}$ genügen der in t quadratischen Gleichung

$$(t-d^2)^2 - (1-d^2)^2 b_j^2 t = o \,.$$

Da in der Gleichung von 7.1 nur d^2 vorkommt, gilt:

7.2 Korollar: Bei $\mathcal{A}$-Matrizen ist das Konvergenzverhalten des S-SOR-Verfahrens nur von $|d|$ abhängig.

Kahan [11] beweist Lemma 7.1, indem er von der Darstellung 5.2 für T_{dd} ausgeht und - ähnlich wie bei der Herleitung von Satz 6.6 - vor allem die bei $\mathcal{A}$-Matrizen geltende Beziehung $L^2 = R^2 = 0$ heranzieht.

7.3 Satz (Kahan): Sei $A = E+B$ eine symmetrische, positiv definite $\mathfrak{A}$-Matrix. Dann gilt:

a) Das S-SOR-Verfahren konvergiert für $-1 < d < 1$.

b) Es ist $d_{oo} = o$.

c) Es ist $\bar{t}_{dd} = \bar{b}^2$ für $d = d_{oo} = o$.

d) Es ist $\bar{t}_{dd} > \bar{b}^2$ für $d \neq o$.

In Worten besagt dieser Satz u.a.: Das S-SOR-Verfahren konvergiert bei symmetrischen, positiv definiten $\mathfrak{A}$-Matrizen für $d = o$ optimal, d.h. die Konvergenz des ursprünglichen Verfahrens von Aitken ($d = o$) läßt sich durch Überrelaxation nicht verbessern.

7.4 Satz: Sei $A = E+B$ eine $\mathfrak{A}$-Matrix und ist B schiefsymmetrisch. Dann gilt:

a) Ist $\bar{b} < 1$, so konvergiert das S-SOR-Verfahren für $-1 < d < 1$.

b) Ist $\bar{b} \geq 1$, so konvergiert das S-SOR-Verfahren für $\sqrt{\frac{\bar{b}-1}{\bar{b}+1}} < |d| < 1$.

c) Es ist $|d_{oo}| = \frac{\bar{b}}{1 + \sqrt{1+\bar{b}^2}}$.

d) Es ist $\bar{t}_{dd} = d^2$ für $|d| \geq |d_{oo}|$.

e) Es ist $\bar{t}_{dd} > d^2$ für $|d| < |d_{oo}|$.

Beweis: Gehen wir von Lemma 7.1 aus und setzen wir

$$d^2 = -u, \tag{7.5}$$

dann erhalten wir die quadratische Gleichung

$$(t+u)^2 - (1+u)^2 b_j^2 t = o,$$

die mit der quadratischen Gleichung von 6.6 identisch ist, falls man d anstelle von u schreibt. Nun war aber 6.6 Ausgangspunkt für den Beweis von 6.8; alle Ergebnisse, die wir dort bezüglich der Größen d_o und d_g ableiteten, können wir nun für entsprechende Größen u_o und u_g übernehmen - wir müssen allerdings berücksichtigen, daß u nach 7.5 nicht positiv sein kann. Aus den Werten u_o und u_g ergeben sich dann die Größen d_{oo} und d_{gg} nach 7.5 .

Wir erhalten aus 6.8b

$$u_o = -\left(\frac{\bar{b}}{1 + \sqrt{1+\bar{b}^2}}\right)^2 = -d_{oo}^2 .$$

Dies ist die Behauptung 7.4c; 7.4d und 7.4e folgen aus 6.8c und 6.8d:

$$\bar{r}_{dd} = |u| = d^2 \quad \text{für } |u| \geq |u_o| \quad \text{bzw. } |d| \geq |d_{oo}|$$

$$\bar{r}_{dd} > |u_o| = d_{oo}^2 \quad \text{für } d \neq d_{oo} .$$

Bezüglich der Größe d_g unterscheiden wir: Ist $\bar{b} \geq 1$, dann ist $u_g = -(\bar{b}-1)/(\bar{b}+1) \leq o$ und $d_{gg} = \sqrt{|u_g|} = \sqrt{(\bar{b}-1)/(\bar{b}+1)}$. Ist $\bar{b} < 1$, dann ist $u_g > o$. Dann konvergiert das S-SOR-Verfahren für alle Werte $-1 < u \leq o$, d.h. für alle d mit $o \leq |d| < 1$. Damit sind auch 7.4a und 7.4b bewiesen.

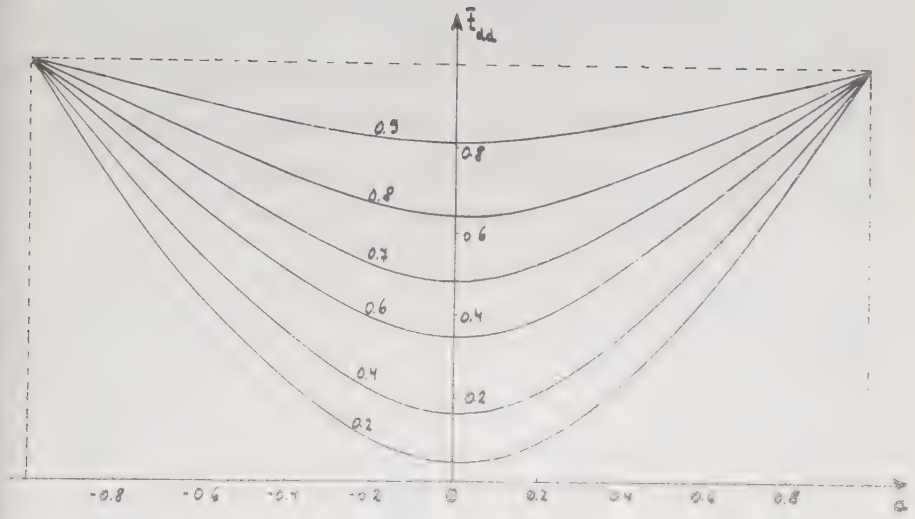
Wir können nun bei einer gegebenen Matrix B die entsprechenden Größen des SOR- und des S-SOR-Verfahrens vergleichen. Es ist

(7.6) $$d_{oo}^2 = |d_o| \, ,$$

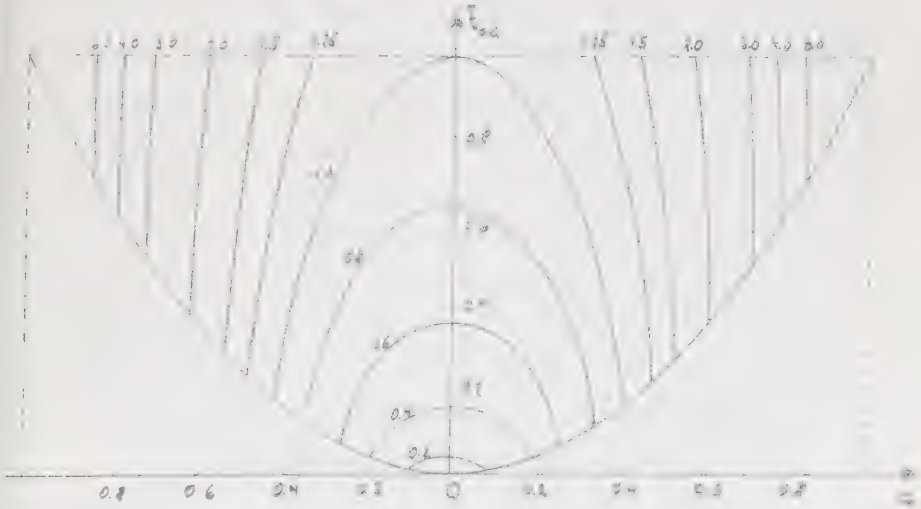
(7.7) $$\bar{t}_{d_{oo}d_{oo}} = d_{oo}^2 = |d_o| = \bar{t}_{d_o} \, ,$$

(7.8) $$d_{gg}^2 = |d_g| \, .$$

7.7 bedeutet, daß bei jeweils optimalem Relaxationsfaktor die Spektralradien von T_{dd} und T_d gleich sind; also werden zur Erzielung einer vorgegebenen Genauigkeit – asymptotisch gesehen – von beiden Verfahren gleich viel Iterationen benötigt. Da aber das S-SOR-Verfahren pro Iteration doppelt so viele Rechenoperationen erfordert wie das SOR-Verfahren, ist das letztere überlegen. Ferner konvergieren SOR- und B-SOR-Verfahren für $d = d_o$ optimal, während das S-SOR-Verfahren – bei dem ein SOR- und ein B-SOR-Schritt in alternierender Reihenfolge ausgeführt werden – nach 7.6 für $d = |d_{oo}| = \sqrt{|d_o|}$ = [illegible] am schnellsten konvergiert. [illegible] zeigt schließlich, daß das S-SOR-Verfahren für $-|d_{gg}|$ [illegible] d [illegible] divergiert, während SOR- und B-SOR-Verfahren für d-Werte aus diesem Intervall noch konvergieren; andererseits konvergiert das S-SOR-Verfahren auch für d_{gg} [illegible] 1, während die beiden anderen Verfahren (für $\bar{b}$ [illegible] 1) für positive Werte von d [illegible] divergieren. Man kann das eben Gesagte auch durch Vergleich der Abbildungen 1 und 2 erkennen; letztere erhält man, indem man nach 7.1 $\bar{t}_{dd}$ als Funktion von d und $\bar{b}$ sowohl für symmetrische wie für schiefsymmetrische Matrizen bestimmt.



a) B symmetrisch



b) B schiefsymmetrisch

Abb. 2: S-SOR-Verfahren bei α-Matrizen A = E+B

$\bar{\tau}_{dd}$ als Funktion von d (Parameter $\bar{b}$)

8. Beispiele und Anwendungen

Wir wollen in diesem Abschnitt ein Beispiel für Satz 3.5 und ein Beispiel für die Sätze 3.11 und 3.15 bringen. Anschließend gehen wir auf eine mögliche Anwendung dieser Sätze bei Fredholmschen Integralgleichungen ein.

8.1 Beispiel 1:

Die Matrix

$$A = \begin{pmatrix} 1 & a & c \\ |a| & 1 & a \\ |c| & |a| & 1 \end{pmatrix}$$

wurde von Collatz [4] gewählt, um das unterschiedliche Konvergenzverhalten des Gesamtschrittverfahrens und das Einzelschrittverfahrens zu zeigen (Ist $A = E+B$, dann konvergiert das Gesamtschrittverfahren, wenn $\bar{b} < 1$ ist). Jedem Punkt (a,c) einer a-c-Ebene entspricht eine Matrix; man kann auf diese Weise in der a-c-Ebene Bereiche angeben,"in denen das Gesamtschritt- bzw. das Einzelschrittverfahren konvergieren". Dasselbe machen wir beim SOR-Verfahren, indem wir für einen vorgegebenen Wert d_g diejenigen Punkte bestimmen, für deren zugehörigen Matrizen der Spektralradius $\bar{t}_{d_g} = \bar{t}_{d_g}(a,c) = 1$ wird. Diese Punkte bilden eine geschlossene Kurve, die den Nullpunkt enthält. Für die zu den inneren Punkten gehörigen Matrizen ist dann das SOR-Verfahren für $-1 < d < d_g$ konvergent.

In Abbildung 3 sind die Kurven für $d_g = 0.8, 0.6, \ldots, -0.6$ eingetragen. Bei der Berechnung geht man vom charakteristischen Polynom von T_d aus,

$$t^3 + t^2(3d^2+(1+d)^3a^2|c| + (1+d)^2(c|c| + 2a|a|) +$$

$$t(3d^2+(1+d)^3a^2c + d(1+d)(c|c| + 2a|a|) + d^3 = 0 ,$$

und sucht für konstante Werte $d = d_g$ und a die Größe c so zu bestimmen, daß die betragsgrößte Wurzel gleich 1 wird; diese erhält man nach dem Bernoulli-Verfahren (mit dem sich in gleicher Weise reelle und komplexe Wurzeln erfassen lassen). Die Rechnungen wurden auf der Tübinger Rechenanlage Siemens 2002 ausgeführt.

Im ersten Quadranten der a-c-Ebene ist A symmetrisch. Nach dem Satz 2.9 von Ostrowski konvergiert das SOR-Verfahren für alle Werte von d in dem Intervall $-1 < d < 1$ genau dann, wenn A positiv definit ist. Es gibt demnach im ersten Quadranten nur eine Randkurve, die die positiv definiten von den übrigen symmetrischen Matrizen trennt.

Im zweiten Quadranten bestimmt c den symmetrischen, a den schiefsymmetrischen Anteil von A ($a \neq 0$, $c \neq 0$) . Der symmetrische Anteil von A ist positiv definit für $0 \leq c < 1$. Die Abbildung 3 zeigt in Übereinstimmung mit Satz 3.5: Für $c < 1$ läßt sich immer ein d_g finden, so daß das SOR Verfahren für $-1 < d < d_g$ konvergiert. Man erkennt auch, daß

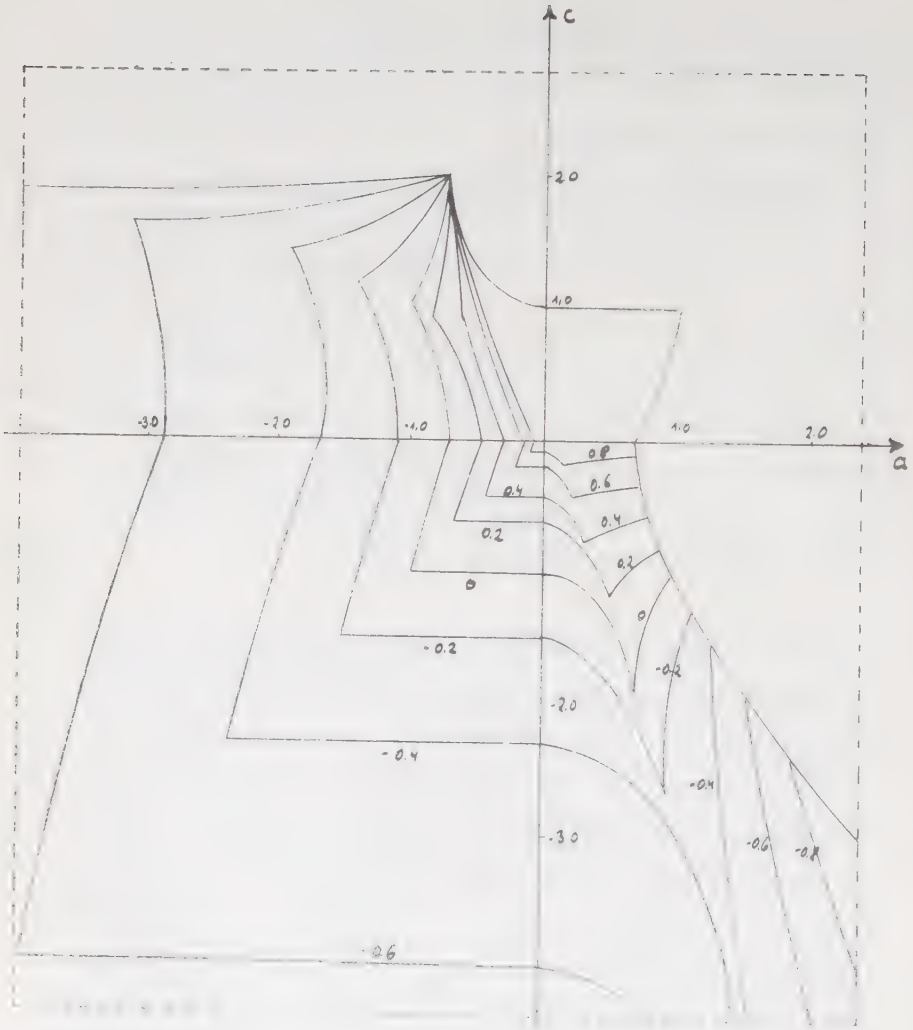


Abb. 3: SOR-Verfahren.

"Konvergenzgebiete" der Matrix $\begin{pmatrix} 1 & a \\ c & 1 \end{pmatrix}$

die Konvergenzbedingung - der symmetrische Anteil von A sei positiv definit - nicht notwendig ist: Es gibt Matrizen mit $c > 1$, für die das SOR-Verfahren konvergiert. Im vierten Quadranten ($a \geq c$, $c \leq 0$) ergibt sich ein bezüglich a und c vertauschtes Bild. Hier ist der symmetrische Anteil von A für $a < 0.5\sqrt{2}$ positiv definit.

Im dritten Quadranten schließlich ist B schiefsymmetrisch. Für jede Matrix läßt sich ein d_g finden, so daß das SOR-Verfahren für $-1 < d < d_g$ konvergiert.

8.2 Beispiel 2:

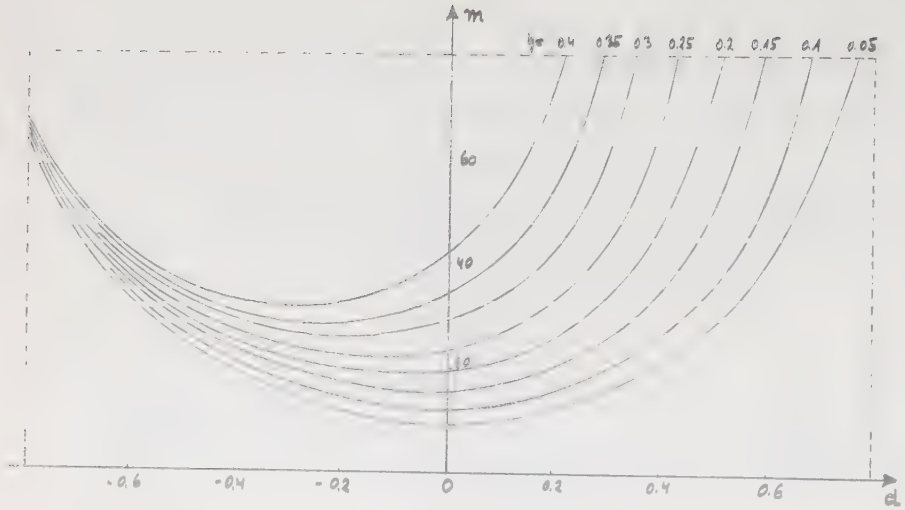
Ist A = E+B und

$$A = \begin{pmatrix} 1 & -h & -h & \cdots \\ h & 1 & -h & \cdots \\ h & h & 1 & \cdots \\ \cdot & \cdot & \cdot & \cdot \end{pmatrix},$$

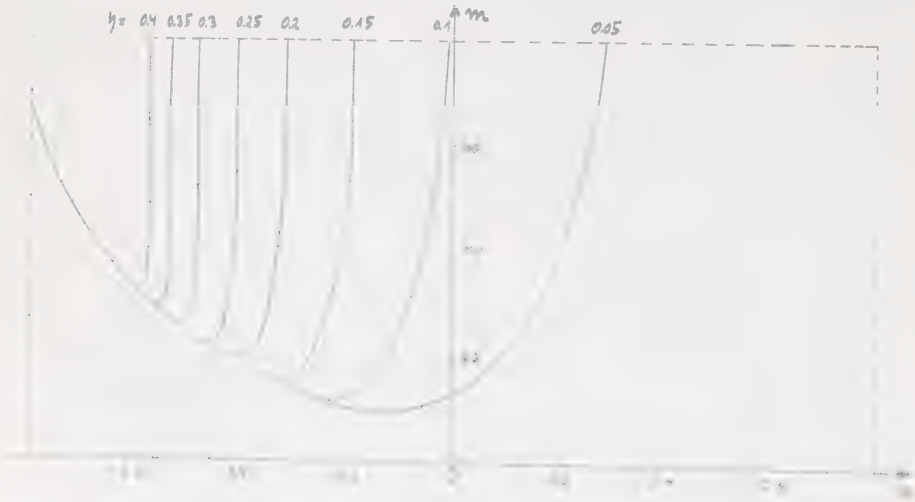
dann ist $B = L-L^T$ schiefsymmetrisch. Die Matrix $R = hE+\tilde{B} = hE+L+L^T$ hat nur Elemente h, hat also den Rang 1. [illegible] Eigenwert von R ist nh, die anderen Eigenwerte sind o. Daraus folgt: $\tilde{B} = R - hE$ hat einen Eigenwert $(n-1)h$ und $n-1$ Eigenwerte h. Für $h > 0$ ist nach Definition 3.2

[illegible]

Nach den Sätzen 3.11 und 3.15 gilt dann für $h > 0$:



a) SOR-Verfahren



b) B-SOR-Verfahren

Abb. 4: Überrelaxation bei der Matrix $A = \begin{pmatrix} 1 & -h & -h & \cdots \\ h & 1 & -h & \cdots \\ h & h & 1 & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{pmatrix}$

Anzahl m der Iterationen als

Funktion von d (Parameter h)

Das SOR-Verfahren konvergiert für $-1 < d < d_g = \frac{1-h}{1+h}$.

Das B-SOR-Verfahren konvergiert für $-1 < d < d_g = \frac{1-(n-1)h}{1+(n-1)h}$.

Um das Ergebnis anschaulicher zu machen, wurden beide Verfahren für Matrizen vom Grad 1o durchgeführt. Ausgangsvektor war $X^{(o)} = 0$. Der Iterationsprozeß wurde jeweils abgebrochen, wenn nach m Iterationen

$$\max_{i=1,..,n} \left| x_i^{(m)} - x_i^{(m-1)} \right| < 1o^{-7}$$

war. In Abbildung 4 ist m in Abhängigkeit von d für verschiedene Werte von h aufgetragen, und zwar in 4a für das SOR-Verfahren, in 4b für das B-SOR-Verfahren (für $h < o$ erhält man gerade das umgekehrte Bild).

8.3 Lineare Fredholmsche Integralgleichungen

Wir wollen hier noch kurz auf eine Anwendung eingehen, bei der z.B. schiefsymmetrische Matrizen auftreten können; es handelt sich um die numerische Lösung linearer Fredholmscher Integralgleichungen 2. Art:[1)]

$$X(t) = C(t) + \int_a^b K(t,s)X(s)ds \qquad (8.3.1)$$

Gegeben sind die stetige Funktion $C(t)$ und der stetige Kern $K(t,s)$ im Intervall $-\infty < a \leq t \leq b < \infty$, $a \leq s \leq b$. Gesucht ist

1) Eine Beschreibung der verschiedenen Lösungsmethoden, unter die sich auch die hier skizzierte Methode einordnet, findet sich bei Walther-Déjon [27].

die Funktion X(t); besitzt die Integralgleichung keinen Eigenwert 1, dann existiert eine im Intervall $a \leq t \leq b$ stetige, eindeutig bestimmte Lösung X(t). Es ist üblich, die Integralgleichung in der Form

(8.3.2) $$X = C - BX$$

zu schreiben. Ersetzt man das Integral nach irgendeiner [illegible]formel durch eine Summe und beschränkt man sich darauf, die gesuchte Funktion nur an den Stützstellen der [illegible]turformel zu bestimmen, dann [illegible] "Integral-operator" B als Matrix schreiben (falls man das Restglied der Quadraturformel vernachlässigt). 8.3.2 ist dann ein [illegible]ares [illegible]system und dessen [illegible] Approximation der gesuchten Funktionswerte in den Stützstellen.

[illegible] system iterativ zu lösen; er dachte vor allem an das Einzel-[illegible] der Kern schiefsymmetrisch ist, d.h. K(t,s) = -K(s,t) ist [illegible] Quadraturformel [illegible] sind, wie dies z.B. bei der Tschebyscheffschen Formel der Fall ist (Vergl. Kantorowitsch-Krylow [12], S.95 ff.). ([illegible] Quadraturformeln erhält man Matrizen B, bei [illegible] Anteil mehr oder weniger über-wiegt.)

Ein schiefsymmetrischer Kern ergibt sich z.B. bei Integralgleichungen vom Faltungstyp $K(t,s) = g(t-s)$, falls g eine ungerade Funktion ist.

Fisher [7] untersucht die Anwendung von Analogrechnern zur Lösung von linearen Integralgleichungen; sein vorgeschlagenes Verfahren der sukzessiven Approximation ist in gewissem Sinn identisch mit dem Einzelschrittverfahren. Der einfache Ansatz, der vom Einzelschrittverfahren zur Überrelaxation führt, läßt sich auch bei Analogrechnern leicht programmieren, so daß man auf diese Weise Konvergenzverbesserungen erzielen könnte.

9. Ergänzungen

9.1: Bemerkungen zur Literatur.

Das Einzelschrittverfahren wurde zum ersten Mal von Gauß in einem Brief erwähnt, in dem die Methode an einem Beispiel beschrieben wird. [8] . Von Seidel stammt die erste ausführliche Beschreibung des Verfahrens, das auch häufig nach Gauß und Seidel benannt wird. Eine Schilderung der historischen Entwicklung findet sich bei Ostrowski [18] und Mises-Geiringer [17] . Verbreitete Anwendung fanden früher allgemeine Relaxationsverfahren, bei denen die Gleichungen nicht in zyklischer Reihenfolge behandelt wurden; die Auswahl der Gleichungen hing dabei weniger von theoretischen Überlegungen, sondern mehr von der praktischen Erfahrung des Rechners ab.

Mit dem Aufkommen programmgesteuerter Rechengeräte tritt die "zyklische Steuerung" in den Vordergrund, weil sich zyklische Verfahren leichter als Algorithmus darstellen lassen. Die Einführung von Überrelaxationsfaktoren geht auf Frankel [7] und Young [29] zurück. Franke. diskutiert die Wahl eines optimalen Relaxationsfaktors für den speziellen Fall der Laplace-Gleichung. Young untersucht $\mathfrak{A}$-Matrizen. Ostrowski [19] führt für eine Matrix vom

Grad 2 eine vollständige Untersuchung des Konvergenzverhaltens durch. Varga [25] und Kjellberg [14] diskutieren eine Verallgemeinerung der Ergebnisse von Young für p-zyklische Matrizen ($\mathcal{A}$-Matrizen sind dann 2-zyklische Matrizen). Für die Praxis ist der Fall p = 2 der wichtigste, weil nur hier symmetrische Matrizen auftreten können. Auch schiefsymmetrische Matrizen kommen nur für p = 2 vor.

Andere allgemeine Überrelaxationsverfahren wie z.B. Blockrelaxation sind in dem Buch von Varga [26] beschrieben. Wir gehen im folgenden noch auf einige mit unseren Untersuchungen zusammenhängende Arbeiten ein.

4.1 Zu den Sätzen von Stein, Weissinger und Ostrowski

Wir formulieren die oben genannten Sätze, die Ausgangspunkt für Abschnitt 3 waren, in der Notation dieses Abschnitts. Sei $A = E+B = E+L+R$, $\tilde{B} = L-R$ und seien $B = G+H$ und $\tilde{B} = \tilde{G}+\tilde{H}$ die Zerlegungen von B bzw. $\tilde{B}$ in symmetrische und schiefsymmetrische Anteile.

Satz (Stein [23]): Sei

$$M = E+\tilde{G}+\frac{1}{2}\left\{H(E+B)^{-1}(E+\tilde{B}) + (E+\tilde{B}^T)(E+B^T)^{-1}H^T\right\}$$

positiv definit. Dann gilt: Das Einzelschrittverfahren konvergiert genau dann, wenn der symmetrische Anteil E+G von A positiv definit ist.

Ein Ergebnis von Weissinger [28] lautet ähnlich.

Ist $K = AA^T$, dann ist $\nu(A) = \sqrt{K}$ die Spektralnorm von A. Zerlegt man nun A = M+N so, daß M positiv definit ist, und ist N = L + R eine Zerlegung von N in eine untere Dreiecksmatrix einschließlich der Diagonalen von N und eine obere Dreiecksmatrix, dann sei $\lambda = \mathrm{Max}(\nu(L), \nu(R))$. Dann gilt:

Satz (Ostrowski [20]): Ist $\lambda <$ [illegible] ,

dann konvergiert das SOR-Verfahren für $-1 < d < 1$.

Die Anwendung der beiden Sätze ist deshalb schwierig, weil sich nicht leicht erkennen läßt, ob die Voraussetzungen erfüllt sind. Grob gesprochen besagen die Sätze, daß die Verfahren konvergieren, wenn die nicht symmetrischen Anteile der Matrix A "genügend klein" sind, d.h. wenn A "fast-symmetrisch" ist (Stein). In Satz 3.5 werden dagegen über den nicht symmetrischen Anteil von A keine einschränkenden Voraussetzungen gemacht; dafür muß man in Kauf nehmen, daß nicht in dem ganzen Intervall $-1 < d < 1$ Konvergenz vorliegt.

9.3 Fehlerabschätzung von Albrecht:

Bei der [illegible] Überrelaxationsverfahren wird man i.a. abbrechen, wenn die Größe von $Z^{(m+1)} = X^{(m+1)} - X^{(m)}$

- gemessen in irgendeiner Norm - unterhalb einer vorgeschriebenen Schranke bleibt. Albrecht [2] zeigt nun für das SOR-Verfahren den Zusammenhang zwischen $Z^{(m+1)}$ und dem eigentlichen Fehlervektor $F^{(m+1)} = X - X^{(m)}$ (X exakte Lösung). Wählt man eine Vektornorm $\| X\|$ und eine damit "verträgliche" Matrixnorm (Vergl. Householder [10]), dann ist

$$\| F^{(m+1)} \| \leq \| A^{-1} \| \cdot \| R+cE \| \cdot \| Z^{(m+1)} \|, \qquad c = \frac{1}{1+d} .$$

Demnach wirkt sich beim SOR-Verfahren - wie bei direkten Verfahren - eine schlechte Kondition ($\| A^{-1} \|$ groß) auf den Fehler aus. Ferner strebt $c \to \infty$ für $d \to -1$, d.h. $\| R+cE \| \to \infty$ für $d \to -1$. Also wird durch einen kleinen Wert von d der Fehler ebenfalls ungünstig beeinflußt.

9.4 Komplexe Überrelaxationsfaktoren.

In einer kürzlich erschienenen Arbeit hat Kredell [15] komplexe Relaxationsfaktoren bei $\mathcal{A}$-Matrizen studiert. Der wichtige Satz 6.6, der die Beziehung zwischen den Eigenwerten von B und den Eigenwerten des Operators T_d herstellt, gilt auch bei komplexen Relaxationsfaktoren; allerdings ist i.a. der Spektralradius $\bar{t}_d$ nicht mehr durch $\bar{b}$ bestimmt.

Ostrowski [20] zeigt, daß das S.R-Verfahren [illegible] hermitesche, positiv [illegible] Matrizen für $0 < \omega < 2$ konvergiert. Ein Beweis von Aitken (er geht von Lemma 3.1 aus) für diesen Satz läßt sich, wie man zeigen kann, auch für komplexe Relaxationsfaktoren $1+d$ verallgemeinern. Da aber komplexe Überrelaxation bei Hermiteschen und erst recht bei reellen Matrizen umfangreichere Rechnungen erfordert, wird man komplexe Relaxationsfaktoren erst dann anwenden, wenn nachgewiesen ist, daß damit eine beträchtliche Konvergenzbeschleunigung zu erzielen ist.

Literaturverzeichnis

1) Aitken, A.C.: On the Iterative Solution of a System of Linear Equations.
Proc. Roy. Soc. Edinb. 63, S. 52-80, (1950).

2) Albrecht, J.: Fehlerabschätzung bei Relaxationsverfahren zur numerischen Auflösung linearer Gleichungen.
Num. Math. 3, S. 188-201, (1961).

3) Carré, B.A.: The Determination of the Optimum Acceleration Factor for Successive Overrelaxation.
Comp. Journ. 4, S. 73-78, (1961).

4) Collatz, L.: Über die Konvergenzkriterien bei Iterationsverfahren für lineare Gleichungssysteme.
Math. Zeitschr. 53, S. 149-161, (1953) .

5) Engeli, M.: Overrelaxation and Related Methods.
Engeli,u.a.: Refined Iterative Methods for Computation of the Solution and the Eigenvalues of Self-adjoint Boundary Value Problems.
Mitt. Nr. 8 Inst. Angew. Math. ETH Zürich.

6) Fisher, M.E.: Proposed Methods for the Analog Solution of Fredholm's Integral Equations.
Journ. Assoc. Comp. Mach. 5, S. 357-369, (1958) .

7) Frankel, S.P.: Convergence Rates of Iterative Treatments of Partial Differential Equations.
Math. Tabl. Aids Comp. 4, S. 65-75, (1950).

8) Gauß, C.F.: Brief an Gerling.
Gauß, C.F., Werke, Band 9, S. 278-281.
W.G. Teubner, Leipzig, (1903).

9) Habetler, G. J.-Wachspress, E. L.: Symmetric Successive Overrelaxation in Solving Diffusion Difference Equations.
Math. of Comp. 15, S. 356-362, (1961).

10) Householder, A.: The Approximate Solution of Matrix Problems.
Journ. Assoc. Comp. Mach. 5, S. 205-243, (1958).

11) Kahan, W.: The Rate of Convergence of the Extrapolated Gauß-Seidel Iterations.
Doctoral Thesis, University of Toronto, (1958).

12) Kantorowitsch, L. W.-Krylow, W. I.: Näherungsmethoden der Höheren Analysis. 611 S.,
Deutscher Verlag der Wissenschaften, Berlin,(1956).

13) Kjellberg, G.: On the Convergence of Successive Overrelaxation applied to a Class of Linear Systems of Equations with Complex Eigenvalues.
Ericsson Technics 14, S. 245-258, (1958).

14) Kjellberg, G.: On the Convergence of Successive Overrelaxation Method for Cyclic Operators.
Num. Math. 3, S. 87-91, (1961).

15) Kredell, B.: On Complex Successive Overrelaxation.
BIT Nord. Tidskr. for Inform. Behandl. 2, S. 143-152, (1962).

16) Kulsrud, H. E.: A Practical Technique for the Determination of the Optimum Relaxation Factor of the Successive Overrelaxation Method.
Comm. Assoc. Comp. Mach. 4, S. 184-187, (1961).

17) Mises, R.v.-Geiringer, H: Praktische Verfahren der Gleichungsauflösung.
Zeitschr. Angew. Math. Mech. 9, S. 59-77 und S. 153-164, (1929).

18) Ostrowski, A.: Determinanten mit überwiegender Hauptdiagonale und die absolute Konvergenz von linearen Iterationsprozessen.
Comment. Math. Helv. 30, S. 175-210, (1953).

19) Ostrowski, A.: On Over- and Under-Relaxation in the Theory of the Cyclic Single Step Iteration.
Math. Tabl. Aids Comp. 7, S. 152-159, (1953).

20) Ostrowski, A.: On the Linear Iteration Processes for Symmetric Matrices.
Rend. di mathematica e delle sue applicazioni. 14, S. 141-163, (1955).

21) Reich, E.: On the Convergence of the Classical Iterative Method for Solving Linear Simultaneous Equations.
Ann. Math. Stat. 20, S. 448-451, (1949).

22) Sheldon, J.W.: On the Numerical Solution of Elliptic Difference Equations.
Math. Tabl. Aids Comp. 9, S. 101-112, (1955).

23) Stein, P.: The Convergence of Seidel Iterants of Nearly Symmetric Matrices.
Math. Tabl. Aids to Comp. 5, S. 237-239, (1951).

24) Stein, P.-Rosenberg, R.L.: On the Solution of Linear Simultaneous Equations by Iteration.
Journ. Lond. Math. Soc. 23, S. 111-118, (1948).

25) Varga, R.S.: P-cyclic Matrices: A Generalization of the Young-Fraenkel Successive Overrelaxation Scheme. Pac. Journ. Math. 9. S. 617-629, (1959).

26) Varga, R.S.: Matrix Iterative Analysis. 322 S. Prentice Hall, Englewood-Cliffs, New Jersey,(1962).

27) Walther, A.-Déjon, B.: General Report on the Numerical Treatment of Integral- and Integro-Differential Equations. Proc. Symp. on Num. Treatm. of Ord. Diff. Equat., Integral and Integro. diff. Equat., Rom 1960, S. 645-671, Birkhäuser Verlag, (1960).

28) Weißinger, J.: Verallgemeinerung des Seidelschen Verfahrens. Zeitschr. Angew. Math. Mech. 33, S. 155-163,(1953).

29) Young, D.: Iterative Methods for Solving Partial Differential Equations of Elliptic Type. Trans. Am. Math. Soc. 76, S. 92-111, (1954).

30) Zurmühl, R.: Matrizen. Springer Verlag, 1960 (zweite Auflage).

Register

(Die Zahlen in Klammern verweisen auf die Seite, auf der die Begriffe eingeführt oder definiert werden.)

$A(=E+B=E+L+R)$: Koeffizientenmatrix des Gleichungssyst. (4)

$B(=A-E)$: (4)

$\mathfrak{A}$-Matrix : (30)

$\bar{b}$, $\tilde{\bar{b}}$: Spektralradien von B bzw. $\tilde{B}$ (5)

B-SOR-Verfahren : (15)

d, 1+d : Relaxationsfaktor (7)

d_g : "Konvergenzgrenze" beim SOR- und B-SOR-Verfahren (11)

d_{gg}: "Konvergenzgrenze" beim S-SOR-Verfahren (41)

Divergenz: (5)

d_o : Optimaler Konvergenzfaktor beim SOR-Verfahren (21)

d_{oo}: Optimaler Konvergenzfaktor beim S-SOR-Verfahren (39)

Einzelschrittverfahren : (4)

G, $\tilde{G}$: Symmetrische Anteile von B bzw. $\tilde{B}$ (11)

H, $\tilde{H}$: Schiefsymmetrische Anteile von B bzw. $\tilde{B}$ (11)

$\bar{h}$, $\tilde{\bar{h}}$: Spektralradien von H bzw. $\tilde{H}$ (11)

Konvergenz : (5)

L: Linke untere Dreiecksmatrix von B (4)

l_G : $= -g_1$ (kleinster Eigenwert der symm. Matrix G) (10)

R: Rechte obere Dreiecksmatrix von B (4)

r_G : $= g_m$ (größter Eigenwert der symm. Matrix G) (10)

SOR-Verfahren : (7)

S-SOR-Verfahren : (23)

Spektralradius : (5)

T_o : Operator des Einzelschrittverfahrens. (5)

T_d : Operator des SOR-Verfahrens (7)

${}_bT_d$: Operator des B-SOR-Verfahrens (16)

T_{dd}: Operator des S-SOR-Verfahrens (24)

$\bar{t}_d$, ${}_b\bar{t}_d$, $\bar{t}_{dd}$: Spektralradien von T_d, ${}_bT_d$, T_{dd} .

Lebenslauf.

Am 2. Dezember 1933 wurde ich, Ernst Wilhelm Niethammer, als Sohn des Landwirts Wilhelm Niethammer und seiner Ehefrau Barbara in Unterjettingen (Württemberg) geboren. Ich besuchte dort die Volksschule und anschließend die Oberschule in Nagold. Nach der Reifeprüfung im Frühjahr 1953 begann ich das Studium der Mathematik und Physik an der Universität in Tübingen. Mit Ausnahme des Sommersemesters 1955 an der Technischen Hochschule in Stuttgart studierte ich immer in Tübingen und legte dort im Frühjahr 1958 die wissenschaftliche Prüfung für das Höhere Lehramt ab. Den anschließenden Vorbereitungsdienst am Seminar für Studienreferendare in Tübingen beendete ich mit dem Assessorexamen im Dezember 1959. Seit Januar 1960 bin ich als wissenschaftlicher Mitarbeiter bei Prof. Dr. Zeller am Rechenzentrum des Mathematischen Institutes in Tübingen angestellt.

Meine akademische Ausbildung verdanke ich den Herren Professoren W. Braunbek, B. Huppert, E. Kamke, H. Kneser, W. Kossel, F. Lösch, G. Möllenstedt, M. Müller, G. Pickert, G. Schulz, H. Wielandt und K. Zeller.

Foto-Druck:
Chemoprint oHG
Gießen